Incorporating Response Times in Item Response Theory Models of Reading Comprehension Fluency

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Shiyang Su

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Mark L. Davison, Dr. Keisha Varma

June, 2017

Acknowledgements

I would like to express my sincere gratitude to many people for their contribution to the journey of my graduate education. First, I would like to give special thanks to my advisor Professor Mark L. Davison for his professional guidance, insight and knowledge. He has taught me many things from writing research proposals to thinking as an independent researcher. Also, thanks to my co-advisor Professor Keisha Varma, without whose support and encouragement I would not have been able to come this far. I would like to thank Professor Michael Rodriguez for his thoughtful feedback and great suggestions during essential stages of building my dissertation. I would also like to thank Professor Gongjun Xu for his professional guidance and kindly help throughout my statistical training.

Likewise, I would like to thank all my friends and colleagues in the Educational Psychology program, from whom I have learned more than any books.

Lastly, I would like to thank my beloved family. With your caring and love I will always have strength.

Dedication

This thesis is dedicated to my family, who are supportive and loving.

Abstract

With the online assessment becoming mainstream and the recording of response times becoming straightforward, the importance of response times as a measure of psychological constructs has been recognized and the literature of modeling times has been growing during the last few decades. Previous studies have tried to formulate models and theories to explain the construct underlying response times, the relationship between response times and response accuracy, and to understand examinees' behaviors.

Different from most existing psychometric models, the current study is based on the idea of reading comprehension fluency in the reading literature and proposes several item response theory based models combining response times and response accuracy. To better understand the construct of reading comprehension fluency, the current study used a new computer-administered assessment of reading comprehension and recorded both the responses and response times of each item. Response times connect examinees' performance on the reading comprehension test to the concepts of fluency or automaticity in the reading literature, concepts that are evidenced by responses that are accurate and appropriately fast. The current study evaluates reading comprehension fluency through two approaches: one with polytomously scored variables and one with conditional variables. The models show the benefits of using the response time information in terms of improving the construct validity when the measured latent construct is reading comprehension fluency. The current study contributes to an interpretation of the latent trait of reading fluency. The models can be used to identify the intervals along the comprehension continuum in which the students tend to read fluently.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

Statement of Problems

With the popularity of online assessments, it becomes inexpensive and common to record response times in psychological and educational testing. The analysis of response times on tests has attracted increasing interest. A question raised is: Does incorporating response time data lead to a better understanding of examinees' test scores? In the past couple decades, researchers tried to formulate models and theories to explain the construct underlying response times as well as the relationship between response accuracy and response times, and to understand examinees' behaviors. Previous studies have used response times as measures of different constructs: for example, in the field of social psychology, response times have been used to measure social desirability (Egloff & Schmukle, 2002; Holden & Kroner, 1992) and attitude strength (Bassili, 1996); in the field of cognitive psychology, response times and response accuracy are the fundamental measures to be considered for cognitive tests. For the purpose of test development, response times can be used to enhance criterion validity and to design better tests. Response times could be used to understand examinees' behaviors. They help differentiate between unreached items and reached items that were not answered, whereas in paper-and-pencil tests such information is difficult to obtain. They allow researchers to evaluate the speededness of a test, and to detect abnormal behaviors such as rapid guessing, cheating, etc. Response times can also be used in cognitive psychology for a more rigorous cognitive theory development (Partchev & De Boeck, 2012).

To use the full diagnostic potentials of response times, psychometric models are needed in order to analyze the relationship between the observed response times and the examinees' latent traits. The estimation of person trait and item parameters can benefit from jointly modeling response times and response accuracy (Molenaar, Tuerlinckx, & van der Maas, 2015; Petscher, Mitchell, & Foorman, 2015; Roskam, 1997; Thissen, 1983; van der Linden, 1999; 2007; Wise & DeMars, 2006). Psychometric models incorporating response time data show practical improvements, such as maximizing the accuracy of person trait estimation and minimizing the standard errors (Petscher et al., 2015; van de Linden, Scrams, & Schnipke, 1999). Therefore, there is a rapid development on how to utilize response times on test items as an additional source of information in estimating examinees' abilities when the test is delivered in a computerized fashion.

However, in contrast to modeling accuracy, there is less agreement on which model(s) to use for response times. How to analyze response times still seems arbitrary to a certain extent. One reason for the disagreement about modeling response times is that there is not much clarity on the nature of the latent trait behind response times. The psychometric literature regards response times and response accuracy as two independent measures and each depends on different constructs. On the other hand, the reading literature relates the response times and response accuracy and concludes they both contribute to the measure of reading comprehension fluency.

To demonstrate the advantage of incorporating response times and to interpret the latent trait measured by response times, the current study used a new computer-

administered assessment of reading comprehension, which records both item responses and item response times. With both measures available, the first goal of the current study is to show that using response times could improve the construct validity when the measured construct is reading comprehension fluency, and to understand the latent trait of reading comprehension fluency. Response times connect examinees' performance on a reading comprehension test to the concepts of fluency and automaticity in the reading literature, concepts that are evidenced by responses that are accurate and appropriately fast (LaBerge & Samuels, 1974). The current study aims at understanding reading comprehension fluency as a construct measured by the product of response accuracy and response times. The models proposed in the current study help us interpret the construct behind fluency -- in other words -- when the respondents answer the item correctly, could they answer this item fast? The usage of response times could facilitate the development of a more rigorous theory of reading comprehension fluency and a better instrument measuring the construct.

The second goal of the current study is to show taking response time information into account could improve the estimation of person trait parameters, using the proposed item response theory (IRT) based models. To address this question, the current study derived IRT based models that capitalize on the response time information while estimating the item parameters and person parameters.

The current study experiments with scores from IRT based models to identify intervals along the comprehension continuum in which students tend to be inefficient or efficient, and to locate students who are able to read fluently. The models are intended to

be practically useful in evaluating test takers' reading comprehension fluency, and helping teachers plan for instructional differentiation.

Overview

As a summary, the dissertation tries to address these questions:

1) How can we best measure reading comprehension fluency? Through the polytomous model or the conditional model?

2) Can we reliably estimate the person trait of comprehension fluency?

3) How much does the trait of comprehension fluency differ from the trait measured by accuracy?

Chapter 2 summarizes the existing studies of response times, and provides a theoretical background of the proposed models. Chapter 3 discusses a new instrument, a reading comprehension test, that is under development, and the evaluation criteria as the dependent variables. Chapter 4 presents the results for two approaches. Chapter 5 concludes the study, further discusses the theoretical framework of reading comprehension fluency, and illustrates the significance and application of the models proposed in the current study.

Chapter II: Literature Review

Introduction

During the past few decades, response times, also called reaction times in cognitive psychology, can be recorded concurrently with the corresponding responses in operational tests. Online assessment is becoming a mainstream form of modern testing due to its flexibility, accessibility, and potential capacities for faster data analysis and reporting. It also makes the collection of examinees' response times more straightforward. Response time is an important concept, which reflects the organization and structure of the cognitive process (Luce, 1986; Ratcliff & Smith, 2004). A growing body of literature suggests that response time is an important factor to consider. Previous studies use response times to measure constructs such as attitude (Bassili, 1996), social desirability (Egloff & Schmukle, 2002; Holden & Kroner, 1992), and to improve criterion-related validity, or to evaluate speededness and detect abnormal behaviors such as cheating, rapid guessing, for purpose of designing a better test (van der Linden & Guo, 2008; van der Linden, 2009).

The accessibility of response times greatly broadens the scope of modeling approaches. There are several advantages of developing models for response times: 1) response times and responses can work together to enhance the prediction of criterion variables, and to improve the criterion-related validity and/or reliability of the measures (e.g., Egloff & Schmukle, 2002; Petscher et al., 2015; Siem, 1996); 2) quite a few studies suggest that response times can be used to improve the estimation of person ability parameters and item parameters (Molenaar et al., 2015; Petscher et al., 2015; Roskam,

1997; Thissen, 1983; van der Linden, 1999; 2007; Wise & DeMars, 2006); 3) modeling response times can help people understand the cognitive processes and provide insights into the cognitive theories (e.g., Partchev & De Boeck, 2012; Scheiblechner, 1979).

A large number of models have been proposed to model response times in the psychometric literature and the cognitive psychology literature. However, there has been a disagreement on how to utilize this information. Models differ in terms of assumptions of the relationship between response times and response accuracy, the nature of the latent traits underlying response times and response accuracy, and the distribution of response times.

Relationship of Response Accuracy and Response Times

Three different approaches have been taken to model response times (Klein Entink, van der Linden, & Fox, 2009; Molenaar et al., 2015; Partchev, De Boeck, & Steyer, 2013). Each approach is reviewed briefly, along with several representative examples. The first approach treats response times and accuracy as two separate constructs and model response times exclusively. For example, Scheiblechner (1979) modeled the response times collected from the speed test with strict time limits. Rouder, Sun, Speckmann, Lu and Zhou (2003) proposed a hierarchical framework to estimate the distribution of response times. Baayen, Davidson, and Bates (2008) and Jaeger (2008) separately proposed a mixed model for response times and a mixed model for response accuracy. Both models introduced the crossed random effects for person and for item when analyzing either response times or response accuracy.

The second approach discusses modeling approaches for response times and

accuracy at the same time, but ignores their correlation. Gorin (2005) regressed the log-transformed response times on the decomposed item difficulty parameters. Mulholland, Pellegrino, & Glaser (1980) used analysis of variance to predict response times by item properties. These models provide some information about both accuracy and response times. However the relation between the variables of response times and response accuracy are unresolved, since such models assume the two variables vary independently. Therefore, a better approach is simultaneous modeling of accuracy and response time as functions of some person and item parameters.

The third approach is then proposed where one can model response times and response accuracy jointly. This approach has been used more frequently in the recent literature. For example, Klein Entink et al. (2009) proposed a model that allows for the simultaneous estimation of ability and speed parameters on the person level and difficulty and time-intensity parameters on the item level. Studies advocating a joint modeling of response time and accuracy include Thissen (1983), Roskam (1997), van der Linden (1999; 2007), Verhelst, Verstralen and Jansen (1997), Wang and Hanson (2005), Klein Entink et al. (2009), Molenaar et al. (2015) etc. Some of these models get ideas from the "speed-accuracy" tradeoff, where an examinee's response accuracy rate might decrease if the examinee chooses to perform a task more quickly (Dennis & Evans, 1996; Luce, 1986). Schnipke and Scrams (1997) used a mixture model to measure speededness, where accuracy depends on examinees' ability, item difficulty and discrimination in a solution behavior. The accuracy rate as a function of response times varies across types of behaviors. Van der Linden (2007) derived a hierarchical framework to model accuracy

exclusively dependent on examinees' ability, and response times exclusively dependent on examinees' latent speeds; in the second level of his model, speed and ability could be correlated. Following this argument, several studies proposed innovative models based on the hierarchical framework, for example, Loeys, Rosseel, and Baten (2011), Wang, Fan, Chang and Douglas (2013), Wang and Xu (2015), etc.

The current study follows the argument of jointly modeling and derives two ways of modeling response times along with response accuracy. A later section will discuss these two approaches.

Modeling Response Times

Following one of these modeling approaches, different models such as IRT models (e.g., Roskam, 1997; Verhelst et al., 1997; Wang & Hanson, 2005), mixed models (e.g., Jaeger, 2008; van Breukelen, 2005), hierarchical structure models (e.g., Rouder et al., 2003; van der Linden, 2007), Cox proportional hazards models (e.g., Ranger & Ortner, 2012; Wang et al., 2013), and mixture models (e.g., Schnipke & Scrams, 1997; Wise & DeMars, 2006) have been investigated in the psychometric literature and the cognitive psychology literature. This section briefly discusses examples of different types of models. Please note that these models can be overlapped, for example, Loeys et al.'s (2011) model employs a hierarchical structure while using mixed effect models on the first level.

IRT Models. Quite a few models have been derived to model response times within an IRT framework. Thissen (1983) was among the first who modeled the response times with an IRT framework. Assuming the random variable of response times follows a

log-normal distribution, the log transformed response time (logtime) of person i to a test item j, $\ln(T_{ij})$, is formulated as

$$\ln(T_{ij}) = \mu + \tau_i + \beta_j - \rho(a_j\theta_i - \beta_j) + \varepsilon_{ij}, \quad (1)$$

where μ is the grand mean of logarithm of response time; τ_i reflects the person slowness parameter for person i; β_j is the item slowness parameter for item j, which can also be interpreted as the amount of time required by item j; a_j, b_j, θ_i reflect the usual discrimination, difficulty, and ability parameters in a two-parameter logistic (2PL) IRT model, and ρ can be interpreted as the regression weight between the ability and the logtime.

Verhelst et al.'s (1997) model assumes the accuracy of an item response depends on the response time, where the probability of getting a correct response increases as the examinee spends more time on the item. The marginal probability of person i answering item j correctly is

$$P\left(X_{ij} = 1 \middle| \theta_i, \tau_i, b_j\right) = 1 - [1 + \exp\left(\theta_i - \ln\tau_i - b_j\right)]^{-\pi_j}, \quad (2)$$

where θ_i is the ability parameter for person i and τ_i is the speed parameter for person i, b_j is the item difficulty parameter for item j, and π_j is the item-dependent shape parameter. When $\pi_j = 1$, this model has the same form as the Rasch (1960) model, with a composite ability parameter $\theta_i - \ln\tau_i$. When the speed parameter τ_i increases, the probability of getting a correct response will decrease given fixed person ability θ_i.

Similarly, a Rasch model was proposed by Roskam (1997) and the probability of person i answering item j correctly is

$$P\left(X_{ij}=1\middle|\theta_i,T_{ij},b_j\right)=\frac{\theta_i T_{ij}}{\theta_i T_{ij}+b_j}=\frac{\exp\left(\xi_i+\tau_{ij}-\sigma_j\right)}{1+\exp\left(\xi_i+\tau_{ij}-\sigma_j\right)}, \quad (3)$$

where θ_i is the ability parameter for person i, T_{ij} is the actual time person i spent on item j, b_j is the difficulty parameter for item j; $\xi_i=\ln\theta_i$, $\tau_{ij}=\ln T_{ij}$, $\sigma_j=\ln b_j$. According to Roskam's (1997) model, when there is more time spent on an item, the probability of answering the item correctly increases; when T_{ij} goes to infinity, the probability of answering the item correct approaches 1 regardless of b_j.

The models derived by Verhelst et al. (1997) and Roskam (1997) assume the examinee's speed is equivalent to the actual response time; their models may not work for adaptive tests where different examinees took different sets of items. The model proposed by Thissen (1983) takes care of this problem by providing distinct parameters for speed and item time intensity.

Wang and Hanson (2005) proposed a four-parameter logistic response time (4PLRT) model for time-limited tests. Their model treats response times as a continuous variable, and assumes that the response times are fixed and independent of the person trait. Adding the response times and item/person slowness parameters to the usual three-parameter logistic (3PL) IRT model, the probability of person i answering item j correct is denoted as

$$P\left(X_{ij}=1\middle|\theta_i,\rho_i,a_j,b_j,c_j,d_j,T_{ij}\right)=c_j+\frac{1-c_j}{1+\exp\left[-1.7a_j\left(\theta_i-\frac{\rho_i d_j}{T_{ij}}-b_j\right)\right]}, \quad (4)$$

where a_j is the discrimination parameter for item j, b_j is the difficulty parameter for item j, c_j is the guessing parameter for item j, and θ_i is the ability parameter for person i, ρ_i is

the person slowness parameter for person i, d_j is the item slowness parameter for item j, and T_{ij} is the response time of person i on item j.

Mixed Models. There have been studies modeling response times with mixed models. For example, a bivariate mixed logistic regression model was chosen by van Breukelen (2005) to predict the log-normalized response times and the log-odds of the correct responses at the same time. Van Breukelen's (2005) model treated person as random but items as fixed.

Assuming item parameters as random variables, a mixed model with crossed random effects for person and items was proposed by Baayen, Davidson, and Bates (2008). Their mixed model for response times is

$$T_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{S}_i\boldsymbol{s}_i + \boldsymbol{W}_j\boldsymbol{w}_j + \varepsilon_{ij}, \quad (5)$$

where T_{ij} is the response time of person i to item j; $\boldsymbol{X}_{ij}\boldsymbol{\beta}$ represents the fixed effects of regression covariates; $\boldsymbol{S}_i\boldsymbol{s}_i$ denotes the random effect of person i; $\boldsymbol{W}_j\boldsymbol{w}_j$ denotes the random effects of item j.

Along the lines of Baayen et al. (2008), a mixed model for response accuracy was presented by Jaeger (2008)

$$logit(p_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{\Theta}_i\boldsymbol{\theta}_i + \boldsymbol{A}_j\boldsymbol{\alpha}_j + \varepsilon_{ij}, \quad (6)$$

where $logit(p_{ij})$ is the log-odds of person i's probability to correctly answer item j; $\boldsymbol{X}_{ij}\boldsymbol{\beta}$ represents the fixed effects of regression covariates; $\boldsymbol{\Theta}_i\boldsymbol{\theta}_i$ denotes the random effect of person i; $\boldsymbol{A}_j\boldsymbol{\alpha}_j$ denotes the random effects of item j. Since the person and item in models of Baayen et al. (2008) and Jaeger (2008) don't have a hierarchical relationship, they are referred to as the crossed random effect models.

Hierarchical Models. Rouder et al. (2003) was among the first to introduce a hierarchical Bayesian framework to estimate the higher-order characteristics of the distribution of response times. Assuming a three-parameter Weibull distribution, the density of response time of person i to item j, T_{ij}, is

$$f(t_{ij}) = \frac{\pi_i (t_{ij} - \psi_i)^{\pi_i - 1}}{\sigma_i^{\pi_i}} \exp[-\left(\frac{t_{ij} - \psi_i}{\sigma_i}\right)^{\pi_i}], t_{ij} > \psi_i, \quad (7)$$

where the Weibull parameters ψ_i, σ_i, π_i are interpreted as the shift parameter described as the lower bound, the scale parameter, and the shape paramter of the distribution. Since this model does not incorporate any item paramters, it essentially regards the response times of person i as distributed identically across items. This model works for the experiemental situations where the cognitive process required by each stimuli is almost the same. To account for variability in both response times and response accuracy, Rouder, Lu, Sun, Speckman, Morey and Naveh-Benjamin (2007) derived another set of hierchical Bayesian models in the context of signal detection.

There were studies suggesting that after controlling for abilities, the fastest examinees were not the most accurate, but fast examinees were consistently fast and slow examinees were consistently slow; in other words, individuals tend to perform at a consistent rate of work across the items, regardless of their ability levels (Kennedy, 1930; Tate, 1948; Schnipke & Scrams, 2002). Following this theory, van der Linden (2007) proposed a hierarchical model, where on the first level response accuracy is an exclusive function of examinee's ability, and response time is an exclusive function of examinee's latent speed, but on the second level the latent speed and ability are allowed to be

correlated. The hierarchical model for person i to item j on the first level includes: 1) a 3PL IRT model for response accuracy with the probability function as

$$P\left(X_{ij}=1\middle|\theta_i,a_j,b_j,c_j\right)=c_j+\left(1-c_j\right)\frac{\exp\left[a_j\left(\theta_i-b_j\right)\right]}{1+\exp\left[a_j\left(\theta_i-b_j\right)\right]}, \tag{8}$$

where a_j, b_j, c_j are the discrimination, difficulty and guessing paramters for item j, and θ_i is the ability parameter for person i; and 2) a lognormal model for response times with the probability function as

$$f\left(t_{ij}\middle|\tau_i,\alpha_j,\beta_j\right)=\frac{\alpha_j}{t_{ij}\sqrt{2\pi}}\exp\{-\frac{1}{2}\left[\alpha_j\left(\ln t_{ij}-\left(\beta_j-\tau_i\right)\right)\right]^2\}, \tag{9}$$

where τ_i is the speed parameter for person i, α_j is the discriminating power for item j, β_j is the time intensity parameter for item j. Also, we know that $\ln t_{ij}\sim N(\beta_j-\tau_i, 1/\alpha_j^2)$.

On the second level, the joint, population-wise distribution of the person parameters is described by a population model, where $\boldsymbol{\xi_i}=(\theta_i,\tau_i)$ is assumed to be randomly drawn from a multivariate normal distribution over P, with mean vector $\boldsymbol{\mu_p}=(\mu_\theta,\mu_\tau)$, and covariance matrix $\boldsymbol{\Sigma_p}=\begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta,\tau}\\ \sigma_{\theta,\tau} & \sigma_\tau^2\end{pmatrix}$. The joint distribution of the item parameters can also be described by a population model, where $\boldsymbol{\psi_j}=(a_j,b_j,c_j,\alpha_j,\beta_j)$ is assumed to follow a multivariate normal distribution over J, with mean vector $\boldsymbol{\mu_J}=\left(\mu_a,\mu_b,\mu_b,\mu_\alpha,\mu_\beta\right)$, and covariance matrix

$$\boldsymbol{\Sigma_J}=\begin{pmatrix}\sigma_a^2 & \sigma_{a,b} & \sigma_{a,c} & \sigma_{a,\alpha} & \sigma_{a,\beta}\\ \sigma_{a,b} & \sigma_b^2 & \sigma_{b,c} & \sigma_{b,\alpha} & \sigma_{b,\beta}\\ \sigma_{a,c} & \sigma_{b,c} & \sigma_c^2 & \sigma_{c,\alpha} & \sigma_{c,\beta}\\ \sigma_{a,\alpha} & \sigma_{b,\alpha} & \sigma_{c,\alpha} & \sigma_\alpha^2 & \sigma_{\alpha,\beta}\\ \sigma_{a,\beta} & \sigma_{b,\beta} & \sigma_{c,\beta} & \sigma_{\alpha,\beta} & \sigma_\beta^2\end{pmatrix}.$$

In fact, the models on the first level can be substituted. Loeys et al.'s (2011) model is an example of employing this hierarchical structure with random effect models on the first level. Another example is that of Wang et al. (2013), who used the Cox proportional hazards model for the response times on the first level.

To incorporate the two mixed models proposed by Baayen et al. (2008) and Jaeger (2008), Loeys et al. (2011) proposed a joint model of response times and accuracy on the second level. The model imposes a multivariate distribution on all random effects for person and item jointly, such that

$$\boldsymbol{\Sigma}_s = \begin{pmatrix} \sigma_\tau^2 & \rho_{\theta\tau}\sigma_\theta\sigma_\tau \\ \rho_{\theta\tau}\sigma_\theta\sigma_\tau & \sigma_\theta^2 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_s = \begin{pmatrix} \sigma_{a1}^2 & \rho_{a1a2}\sigma_{a1}\sigma_{a2} \\ \rho_{a1a2}\sigma_{a1}\sigma_{a2} & \sigma_{a2}^2 \end{pmatrix},$$

where $\rho_{\theta\tau}$ is the correlation between speed and ability at the person level, and ρ_{a1a2} is the correlation between time intensity and difficulty at the item level. This model is built upon van der Linden's (2007) hierarchical framework. It treats the person and items as random and allows for correlation between responses times and accuracy.

Cox Proportional Hazards Models. Survival models such as the Cox proportional hazards model have been used to model response times (Ranger & Ortner, 2011). Wang et al. (2013) embedded the Cox proportional hazards model with a latent speed covariate to model the response times within van der Linden's (2007) hierarchical framework. The Cox proportional hazards model on the first level is

$$h_{ij}(t|\tau_i) = h_{0j}(t)\exp(\beta_j\tau_i), \tag{10}$$

and the survival function is

$$S_{ij}(t) = P\left(t_{ij} > t \middle| \tau_i\right) = \exp\left[-\int_0^t h_{0j}(s)\exp\left(\beta_j\tau_i\right)ds\right], \quad (11)$$

where τ_i is the speed parameter for person i, β_j is the regression weight for item j, and $h_{0j}(t)$ is the baseline hazard which reflects the flexibility to accommodate a variety of different shapes of response time distributions for different items.

Mixture Models. The estimation of item and person parameters in the IRT models could be greatly biased by the speededness. Most IRT models assume that an examinee seeks to answer an item correctly by carefully considering each part of the item; therefore, the probability of obtaining a correct response increases monotonically as the examinee's ability increases. However, if the examinee answers the item quickly without processing the meaning of the question, the correct response probability becomes independent of the ability, and the assumption of IRT models fails in this situation (Wise & Kong, 2005). The two types of behaviors that an examine might employ during a test are named as the solution behavior and the rapid guessing behavior respectively (Schnipke & Scrams, 1997; Wise & DeMars, 2006). Under the solution behavior, the probability of answering an item correctly will be a monotonically increasing function of the examinee's proficiency. Under the rapid-guessing behavior, the probability of a correct response is a constant regardless of the examinee's proficiency. Response times will provide information to better distinguish these two types of behaviors.

Two-state mixture models (Luce, 1986; Townsend & Ashby, 1983) were applied in some studies to account for the effect of the rapid guessing behavior. A HYBRID model of the IRT models and latent class model (Yamamoto, 1989) was first developed and

applied for the examinees with either the solution behavior or the rapid guessing behavior. This model essentially assumes the rapid-guessing behavior happens only toward the end of the test and it doesn't allow for switching back and forth between the two types of behaviors, whereas in reality the rapid guessing behavior can be found throughout the test.

Schnipke and Scrams (1997) used a two-state mixture model to measure speededness. They inspected where the distributions of the rapid-guessing and the solution behavior crossed for each item and used the cross points as the thresholds of response times. Their two-state mixture model decomposed the distribution of observed response times for each item into two weighted components: one for response times from the solution behavior and one for response times from the rapid guessing behavior. Their equation is

$$F_j(\theta) = \rho_j F_{sj} + (1 - \rho_j) F_{gj}, \quad (12)$$

where F_{sj} is the distribution of the responses times for solution behaviors for item j, F_{gj} is the distribution of the response times for rapid guessing behaviors for item j, ρ_j is the proportion of solution behaviors of item j. This model assumes an increasing proportion of rapid-guessing behaviors toward the end of the test.

Wise and Kong (2005) developed an index called response time effort (RTE) to measure an examinee's overall test taking effort using the proportion of test items for which the examinee exhibited solution behaviors. The time thresholds distinguishing two behaviors were identified by visually inspecting the response time frequency distribution. Wise and DeMars (2006) developed an IRT model that incorporated examinees' effort,

represented by response times, and to evaluate the model when the rapid-guessing behavior was present. Suppose the solution behavior is represented by a 3PL IRT model and the rapid-guessing behavior is represented by a constant probability function g_j, then their effort-moderated model is

$$P\left(X_{ij}=1\middle|\theta_i,a_j,b_j,c_j\right)=SB_{ij}\left\{c_j+\left(1-c_j\right)\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}\right\}+(1-SB_{ij})g_j, \quad (13)$$

where a_j, b_j, c_j are the discrimination, difficulty and guessing paramters for item j and θ_i is the ability parameter for person i; g_j equals the reciprocal of the number of response options for item j; SB_{ij} is a dichotomous index indicating a solution behavior of person i on item j, and is treated as known; $SB_{ij}=1$ indicates that person i's response time on item j is from a solution behavior. The maximum log-likelihood function of parameter estimation works the same as the usual IRT model, since the rapid guessing function is a constant. Their study of real data with a sample of sophomores showed that, compared with the usual 3PL model, the effort-moderated model fitted the examinees' response patterns better, provided different but more accurate item parameter estimates, had a lower information function and a lower reliability for the examinees' displaying no rapid-guessing behavior, and showed higher convergent validity through correlations with the external variables including SAT-Verbal, SAT-Quantitative and GPA.

Wang and Xu (2015) argued that Wise and DeMars's (2006) model was not strictly a mixture model since SB_{ij} was determined in advance based on the response time distribution. Instead, they proposed a flexible mixture hierarchical model to account for the differences within response times and response accuracy arising from solution

behaviors and rapid guessing behaviors. They used a latent variable, Δ_{ij}, as a binary indicator variable indicating whether person i responsed to item j using a solution behavior. Following van der Linden's (2007) hierarchical framework, on the first level they proposed mixture models for response times and for response accuracy respectively. Their model also serves to identify the specific type of behavior an examinee employs on an item.

Transformation and Distribution of Response Times

Besides the modeling mechanism and its relationship with response accuracy, another issue to consider for response times is the assumed distribution (e.g., lognormal, exponential, gamma, Weibull, etc.). Since the variable of response times is non-negative, a distribution defined on the entire real continuum is not appropriate for response times (van der Linden, 2006). Misspecification of the distribution of response times might cause invalid inferences.

The lognormal distribution has been widely assumed when modeling response times (e.g., Thissen, 1983; van der Linden, 2007). For example, Thissen's (1983) model assumed the random error term ε_{ij} of the logtime $\ln T_{ij}$ followed a normal distribution, which implied that the model belonged to a lognormal family. Even though the assumption that the response times are distributed log-normally is common in response time models, other distributions such as the Weibull, gamma, and exponential have also been proposed for modeling response times of test items.

Several studies assume the random variable of times spent on an item, T, follows a gamma distribution (Verhelst et al., 1997) with the probability density function as

$$g(t) = \frac{\partial P(T \leq t)}{\partial t} = \frac{\beta^p}{\Gamma(p)} t^{p-1} \exp(-\beta t), \quad t \geq 0, p > 0, \beta > 0. \tag{14}$$

If the expectation of this random variable is $E[t] = p/\beta$, then β/p is the speed parameter, which represents the expected number of finished items per unit time. Maris (1993) formulated a more general gamma distribution for response times with the additive, multiplicative, and combined additive-multiplicative models.

The exponential distribution was chosen in Scheiblechner's (1979) model, where the random variable of response time of person i to item j has the probability density function as

$$f(t_{ij}) = (\tau_i + \gamma_j) \exp[-(\tau_i + \gamma_j) t_{ij}], \tag{15}$$

where τ_i is the person speed parameter; γ_j is the item speed parameter and can be further decomposed into

$$\gamma_j = \sum_{k=1}^{K} \alpha_{jk} \eta_k + c, \tag{16}$$

where α_{jk} is the weight specifying the experimental design or the psychological model; η_k is the time intensity parameter of component k; c is the normalization constant.

The Weibull distribution has also been applied in some studies (e.g., Loeys et al., 2011; Rouder et al., 2003). There are three properties of response times that should be considered when specifying its distribution: 1) participant and item variability; 2) increasing variance with means, 3) a non-zero minimum value (Rouder, Tuerlinckx, Speckman, Lu & Gomez, 2008). A log-transformation of response times could satisfy the first two properties, but not the last property. Rouder et al. (2003; 2008) showed a shifted three-parameter Weibull distribution could satisfy the three properties.

On condition that the lognormal transformation might violate the normality assumption, a broader class of Box-Cox transformations were consider to model response times. With the Box-Cox normal model (Klein Entink et al., 2009), a power parameter was used to represent the transformation. Note that the lognormal distribution is a special case of the Box-Cox transformation.

Except for deciding a proper distribution for response times, Ranger and Kuhn (2012) indicated that the distribution of response times differed dramatically across items within a test. Therefore, a flexible model relaxing the assumption of the response time distribution is usually preferred.

From the discussion above, we can see that in contrast to modeling response accuracy, there is much less agreement on what approach and which model to use for response times. Despites the number of modeling approaches and the candidate distributions to choose from, how to model the response times still seems arbitrary to certain extent.

Latent Trait behind Response Times

One reason for the disagreement about modeling response times is that there is not much clarity on the nature of the latent trait behind the response times. This section discusses the concepts of response times and speed, the relationship between speed and accuracy, and the construct of fluency and automaticity.

Response Times and Speed. The concepts of ability (i.e., level, power) and speed were proposed by Thorndike, Bregman, Cobb, & Woodyard (1926) and were

empirically defined by the produced products, which are the item responses, and the required time to produce products, which are the response times. A variety of studies have been conducted to conceptualize the latent trait underlying the observed responses and response times, in the field of psychometrics (e.g., van der Linden, 2007; Wang & Hanson, 2005) and the field of cognitive psychology and information processing (e.g., Partchev & De Boeck, 2012; Ratcliff & Smith, 2004). According to van der Linden (2009), response time and speed are two different concepts, but are closely related. The response time of an item can be decomposed into two parameters: one for the speed of the person and one for the time intensity of the item. This can be written as

$$t_{ij} = \frac{\beta_j^*}{\tau_i^*} , \tag{17}$$

where t_{ij} is the response time of person i to item j; β_j^* is the time intensity parameter of item j; τ_i^* is the speed parameter of person i. Controlling for the time intensity of items across the test, response time becomes a linear function of the person's speed.

Speed and Ability. Ability and speed are defined as the fundamental concepts to be considered for cognitive tests (Thorndike et al., 1926). The relation between speed and ability is controversial. Earlier in psychology, speed and level were assumed dependent on the same ability (Spearman, 1927); later on, studies concluded that speed and accuracy were distinct since they loaded on different factors (Davidson & Carroll, 1945). Some psychometric literature regards response times and accuracy as two independent constructs, with a very low or negative correlation (van der Linden et al., 1999; van der Linden, 2009). Hambleton and Swaninathan (1985) argued that speed and ability components would require separate dimensions. Partchev, De Boeck and Steyer (2013)

used a verbal analogy test with time limits and found that speed and ability were more or less uncorrelated.

On the other hand, some studies suggest that ability and speed jointly affect response behaviors in assessment instruments (e.g., Lohman, 1989; van der Linden, 2009). The speed-accuracy tradeoff has been discussed in studies of experimental psychology and cognitive psychology. The tradeoff suggests an examinee's response accuracy rate might decrease if the examinee chooses to perform a task quicker (Dennis & Evans, 1996; Kahane & Loftus, 1999). According to Schnipke and Scrams (1997), the accuracy rate as a function of response time doesn't seem to be linear: accuracy levels are low with short response times, and rise to higher levels with longer response times, and reach a plateau when adding response times don't increase accuracy. Also, the correlation between accuracy and speed differs depending on the test content and context (Schnipke & Scrams, 2002).

The current study considers a joint perspective of speed and accuracy and how they describe the examinees' behaviors in the reading comprehension test.

Fluency and Automaticity. Reading comprehension fluency, also called comprehension automaticity, is an important skill for readers to develop. The reading literature relates the accuracy and speed and defines their composite as fluency or automaticity (Chard, Vaughn, & Tyler, 2002; LaBerge & Samuels, 1974). In LaBerge and Samuels's (1974) theory of automatic information processing, they indicated that learning to read involves increasing automaticity in processing word unites into recognizable words and connecting the words when reading a passage. The reader's

understanding of the meaning of the text is based upon the cognitive processing of units, words and connected text. Therefore, reading comprehension fluency or comprehension automaticity is crucial for the development of reading comprehension. Students with reading or learning disabilities are more at risk of difficulties in reading comprehension fluency (Meyer & Felton, 1999). Speed plays a prominent role in improving the cognitive processing. Perfetti (1985) suggested that slow word processing speed interferes with reading comprehension fluency by consuming the working memory with words, and therefore, preventing the reader from thinking about the content while reading. The current study denotes the speed of choosing a correct response as comprehension efficiency. Different from comprehension accuracy, comprehension efficiency is defined as the rate at which one can arrive at a correct response, and it reflects the speed to achieve accuracy.

Previous studies implied that fast responses are processed differently than slow responses. Information processing is based on two fundamental modes: controlled and automatic. It is difficult to suppress or to alter an automatic process once learned. Specifically, fast responses require more automatic processing whereas slow responses require more controlled processing (Goldhammer, Naumann, Stelter, Tóth, Rölke, & Klieme, 2014; Petscher et al., 2015; Shiffrin & Schneider, 1977). Partchev and De Boeck (2012) indicated that fast correct responses and slow correct responses involved different processes and abilities in a matrices test and a verbal analogies test. However, their model started from the differentiation between fast and slow and then made further differentiation between correct and incorrect conditional on the speed; therefore, their

study was not about the construct of fluency. Instead, the current study starts from the model of accuracy and then further differentiated it by speed. In an effort to measure whether there are different cognitive processes behind the same accurate responses in a reading comprehension test, the current study focuses on providing information about comprehension efficiency. Based on the reading literature, the current study denotes the comprehension accuracy and comprehension efficiency as indicators of reading comprehension fluency.

In most of the previous studies in fields of psychometrics or cognitive psychology, the response time data has been treated as a continuous variable. On the contrary, Partchev et al. (2013) treated the response time as a dichotomous variable by categorizing the item response times into fast and slow using the empirical median as the cut off. Similarly, instead of treating response times as continuous variables, the current model dichotomizes the item response times based on whether or not the response is reached in a time interval above or below the cut-off – median in this case. There are several advantages of using a median split of response times. First of all, as I discussed before, even though there exist various models for response times, there is little agreement on which ones to use. By dichotomizing the response times, the current study is able to utilize the usual IRT approach and avoids the complexity of choosing from the various existing models of response times. Second, there have been debates regarding what distribution the random variable of response times follows. As discussed above, distributions such as lognormal (van der Linden, 2006; 2007), gamma (Maris, 1993; Verhelst et al., 1997), Weibull (Rouder et al., 2003) were used to model response times;

transformations on response time variables such as logarithm or Box-Cox transformation (Klein Entink et al., 2009) were applied in previous studies. In the current study, by using the median split to dichotomize response times, a response above beyond the median becomes invariant with respect to the form of the distribution or a monotone transformation of response times; in other words, respondents who are categorized as fast regarding the median will not change as a function of the chosen distribution or the transformation method. Third, since a respondent who guesses tends to proceed rapidly, approaches treating the response time as continuous variables might end up giving more credit than necessary to those fast guessers and therefore, bias our estimation of person ability parameters. To avoid this, the current model scores the response time data into dichotomous variables and gives the same credit to fast guessers and fast-and-accurate respondents in terms of speed. One disadvantage of treating response times as a dichotomous variable is that one would end up losing information; however, given the benefits I discussed, the current study dichotomizes the item response time data using a median split, and utilizes the IRT based models which do not require a lot of computational power.

Polytomous Models and Conditional Models

The current study connects response times with response accuracy to the concepts of fluency and automaticity in the reading literature, which refers to a response that is accurate and appropriately fast. The current study intends to investigate the construct behind fluency, in other words, when the respondents answer the item correctly, could

they answer this item fast? If the examinee answers an item correctly and fast, then I assume it is an automatic process for the examinee to understand the meaning of the item; if the examinee answers an item correctly but slow, then I assume the automatic process of reading comprehension is not fully developed yet.

Unidimensional Polytomous Models. To understand the latent trait measured in the proposed reading assessment, I proposed two approaches to model the reading comprehension fluency. The first approach involves different polytomous models. By incorporating the response time categories, the dichotomous responses of accuracy are transferred into polytomous responses of reading comprehension fluency. The polytomous scoring of an item provides more information than the dichotomous scoring of the same item (Samejima, 1969). The section below discusses several unidimensional models for polytomously scored items, with either ordinal categories or nominal categories.

Partial Credit Model. One popular parametric IRT model for the polytomous items with ordinal categories is the partial credit model (PCM; Masters, 1982). It was originally developed for analyzing test items that require multiple steps and for which it is important to assign partial credit for completing several steps in the solution process. The PCM is a divide-by-total IRT model and can be considered as an extension of the Rasch Model.

Let $X_j = l; l = 1, \dots, L$, then the probability of an examinee with ability θ getting to category x of item j is denoted as

$$P_{jl}(\theta) = \frac{e^{[\sum_0^x(\theta-\delta_{jl})]}}{\sum_0^m e^{[\sum_0^r(\theta-\delta_{jl})]}}, \tag{18}$$

where item j is scored $x = 0, \ldots, m$, with $L = m + 1$ categories; δ_{jl} is the difficulty parameter of step l on item j. For the step 0, the equation is written as

$$P_{jl}(\theta) = \frac{1}{\sum_0^m \mathrm{e}^{[\sum_0^r(\theta - \delta_{jl})]}}. \tag{19}$$

General Partial Credit Model. If the item-level discrimination parameter is included and is held constant across the item steps of the same item, the generalized partial credit model (GPCM) is obtained (Muraki, 1992), and the probability function in equation (18) becomes

$$P_{jl}(\theta) = \frac{\mathrm{e}^{[\sum_0^x a_j(\theta - \delta_{jl})]}}{\sum_0^m \mathrm{e}^{[\sum_0^r a_j(\theta - \delta_{jl})]}}, \tag{20}$$

where a_j is the discrimination parameter common across all steps, but unique to item j. Fixing $a_j = 1$ across items the equation reduces to the PCM. The GPCM allows the possibility of identifying item response options that may be redundant with each other.

Graded Response Model. Another parametric IRT model for the polytomous items with ordinal categories is the graded response model (GRM) (Samejima, 1969). Let $X_j = l; l = 1, \ldots, L$, and define a_j as the item discrimination parameter for item j, λ_{jl} as the threshold parameter for item j in category l. The logistic form of the unidimensional GRM is:

$$P_{jl}^{+}(\theta) = P\left(X_j \geq l \middle| \theta\right) = \frac{\mathrm{e}^{[a_j(\theta - \lambda_{jl})]}}{1 + \mathrm{e}^{[a_j(\theta - \lambda_{jl})]}}, \tag{21}$$

where $P_{jl}^{+}(\theta)$ represents the probability of an examinee with ability θ choosing category l or a higher category for item j. Similarly, the probability of an examinee with ability θ choosing category $l + 1$ or a higher category for item j is denoted as

$$P_{j,l+1}^{+}(\theta) = P\left(X_j \geq l+1 \middle| \theta\right) = \frac{e^{[a_j(\theta-\lambda_{j,l+1})]}}{1+e^{[a_j(\theta-\lambda_{j,l+1})]}}. \quad (22)$$

Note that $P\left(X_j \geq 0 \middle| \theta\right) = 1$ and $P\left(X_j \geq L+1 \middle| \theta\right) = 0$. The probability of choosing category l for item j is then

$$P_{jl}(\theta) = P_{jl}^{+}(\theta) - P_{j,l+1}^{+}(\theta) = \frac{e^{[a_j(\theta-b_{jl})]}}{1+e^{[a_j(\theta-b_{jl})]}} - \frac{e^{[a_j(\theta-b_{j,l+1})]}}{1+e^{[a_j(\theta-b_{j,l+1})]}}, \quad (23)$$

where $P_{jl}(\theta)$ is the probability of a randomly selected examinee with latent trait θ for item j in category l.

Nominal Response Model. In the situation where item response options are not necessarily ordered, the nominal response model (NRM; Bock, 1972) can be used. For an examinee with ability θ, the item response function for the NRM is

$$P_{jl}(\theta) = \frac{e^{[a_{jl}\theta+c_{jl}]}}{1+\sum_{l=1}^{m} e^{[a_{jl}\theta+c_{jl}]}}, \quad (24)$$

where a_{jl} is the discrimination parameter and c_{jl} is the location parameter. The NRM has been proposed to account for guessing behaviors for examinees with low ability in a reading comprehension test (e.g., Thissen, Steinberg, & Mooney, 1989).

Conditional Response Models. The second approach is a conditional approach. A concept named the response contingent item response theory (RCIRT) is introduced (Davison et al., 2016). In the RCIRT, the model of a response to an item is contingent on responses to one or more earlier items, for example, whether item j's response variable X_j satisfies a specified model is contingent on one or more other observed responses. X_j is the dependent item response, which is called the child variable; the item response

variable(s) on which it is dependent on, X_{kj}, is called the parent(s) k of item j. In the current study, both the parent and child variables are dichotomous variables, however, the RCIRT can be readily generalized to polytomous variables. Here I discuss two examples in the literature that fit the broad concept of the RCIRT models.

Andrich & Kreiner (2010). Andrich and Kreiner (2010) were concerned about the violation of local independence when the response to one item governs the response to a subsequent item. In their example items on two adjacent math problems might violate the assumption of local independence, since answering the later item (child) presumes answering the earlier item (parent) correctly. Using a Rasch (1960) model, the response dependence model can be formulated as a parent model and two child models, each conditional on a response of the parent. The probability of people answering the parent item kj correctly is

$$p_{kj} = P\left(X_{kj} = 1 \middle| \theta, b_{kj}\right) = \frac{\exp\left(\theta - b_{kj}\right)}{1 + \exp\left(\theta - b_{kj}\right)}, \tag{25}$$

where θ is the math ability parameter, and b_{kj} is the difficulty parameter of the parent item. The probability of answering the child item j correctly when people answered the parent item kj correctly is

$$p_{j1} = P\left(X_j = 1 \middle| X_{kj} = 1, \theta, b_{j1}\right) = \frac{\exp\left(\theta - b_{j1}\right)}{1 + \exp\left(\theta - b_{j1}\right)}, \tag{26}$$

where θ is the math ability parameter, and b_{j1} is the difficulty of item j for people who correctly answered parent item kj. The probability of answering the child item j correctly when people answered the parent item kj incorrectly is

$$p_{j0} = P\left(X_j = 1 \middle| X_{kj} = 0, \theta, b_{j0}\right) = \frac{\exp\left(\theta - b_{j0}\right)}{1 + \exp\left(\theta - b_{j0}\right)}, \tag{27}$$

where θ is the math ability parameter, and b_{j0} is the difficulty parameter of item j for people who incorrectly answered parent item kj. These two child models differ on the difficulty parameters, but not on the ability parameter.

Since the items are not independent any more, the assumption of local independence is violated. Therefore, the formulation of the likelihood function involves some conditional probabilities, which means the probabilities of the response variables are conditional on other observed responses. Given a value of ability θ in the model of Andrich and Kreiner (2010), the likelihood of a response pattern for the parent and child items can be expressed as (Davison et al., 2016)

$$\begin{aligned} L\left(X_{kj}, X_j\right) &= {p_{kj}}^{x_{kj}}(1 - p_{kj})^{1-x_{kj}}[{p_{j1}}^{x_j}(1 - p_{j1})^{1-x_j}]^{x_{kj}}[{p_{j0}}^{x_j}(1 - p_{j0})^{1-x_j}]^{1-x_{kj}} \\ &= {p_{kj}}^{x_{kj}}(1 - p_{kj})^{1-x_{kj}}{p_{j1}}^{x_j x_{kj}}(1 - p_{j1})^{(1-x_j)x_{kj}}{p_{j0}}^{x_j(1-x_{kj})}(1 - p_{j0})^{(1-x_j)(1-x_{kj})}. \end{aligned} \tag{28}$$

This likelihood function contains the conditional probabilities, which have the parent variables as an exponent.

To estimate the parameters, the response vector (X_{kj}, X_j) could be replaced with multiple dummy variables, each reflecting a conditional item response model. The child variable x_j could be replaced with two dummy variables Y_{j1} and Y_{j0}:

$$Y_{j1} = \begin{cases} X_j, & \text{if } X_{kj} = 1; \\ missing, & \text{if } X_{kj} = 0. \end{cases} \tag{29}$$

$$Y_{j0} = \begin{cases} X_j, & \text{if } X_{kj} = 0; \\ missing, & \text{if } X_{kj} = 1. \end{cases} \tag{30}$$

Therefore, the child variable X_j could be replaced by Y_{j1} if the equation is conditional on $X_{kj} = 1$, and Y_{j0} if the equation is conditional on $X_{kj} = 0$. Equation (26) then becomes

$$p_{j1} = P\left(Y_{j1} = 1 \middle| X_{kj} = 1, \theta, b_{j1}\right) = \frac{\exp\left(\theta - b_{j1}\right)}{1 + \exp\left(\theta - b_{j1}\right)}, \tag{31}$$

and Equation (27) becomes

$$p_{j0} = P\left(Y_{j0} = 1 \middle| X_{kj} = 0, \theta, b_{j0}\right) = \frac{\exp\left(\theta - b_{j0}\right)}{1 + \exp\left(\theta - b_{j0}\right)}. \tag{32}$$

The likelihood function in Equation (28) can be rewritten in terms of the response vector (X_{kj}, Y_{j1}, Y_{j0}) as

$$L\left(X_{kj}, Y_{j1}, Y_{j0}\right) = {p_{kj}}^{x_{kj}}(1 - p_{kj})^{1-x_{kj}}{p_{j1}}^{y_{j1}}(1 - p_{j1})^{1-y_{j1}}{p_{j0}}^{y_{j0}}(1 - p_{j0})^{1-y_{j0}}. \tag{33}$$

This is because that, when $x_j = 1$ and $x_{kj} = 1$, $x_j x_{kj} = y_{j1} = 1$; otherwise, $x_j x_{kj} = 0$, and $y_{j1} = 0$ or missing. When $x_j = 0$ and $x_{kj} = 1$, $\left(1 - x_j\right)x_{kj} = 1 - y_{j1} = 1$; otherwise, $\left(1 - x_j\right)x_{kj} = 0$, and $1 - y_{j1} = 0$ or missing. When $x_j = 1$ and $x_{kj} = 0$, $x_j(1 - x_{kj}) = y_{j0} = 1$; otherwise, $x_j(1 - x_{kj}) = 0$, and $y_{j0} = 0$ or missing. When $x_j = 0$ and $x_{kj} = 0$, $\left(1 - x_j\right)\left(1 - x_{kj}\right) = 1 - y_{j0} = 1$; otherwise, $(1 - x_j)(1 - x_{kj}) = 0$, and $1 - y_{j0} = 0$ or missing.

The Fisher information functions of the child variable X_j and two dummy variable Y_{j1}, Y_{j0} are functions of the same latent trait θ. The information function of the child variable X_j varies depending on the value of the parent variable. When the response to the parent item is $X_{kj} = 1$, the probability function of the child variable X_j is $p_{j1} = P_{j1}(\theta)$, and its information function is

$$I_1(X_j|\theta) = \frac{(p'_{j1})^2}{p_{j1}(1-p_{j1})} \tag{34}$$

where $p'_{j1} = \frac{\partial P_{j1}(\theta)}{\partial \theta}$. When the response to the parent item is $X_{kj} = 0$, the probability function of the child variable X_j is $p_{j0} = P_{j0}(\theta)$, and its information function is

$$I_0(X_j|\theta) = \frac{(p'_{j0})^2}{p_{j0}(1-p_{j0})} \tag{35}$$

where $p'_{j0} = \frac{\partial P_{j0}(\theta)}{\partial \theta}$. Thus the expected information function of the child variable X_j is

$$E[I(X_j|\theta)] = p_{kj}\frac{(p'_{j1})^2}{p_{j1}(1-p_{j1})} + (1-p_{kj})\frac{(p'_{j0})^2}{p_{j0}(1-p_{j0})}\ , \tag{36}$$

where p_{kj} is the probability when $X_{kj} = 1$, and $1 - p_{kj}$ is the probability when $X_{kj} = 0$.

Similarly, the information functions for the dummy variables Y_{j1}, Y_{j0} vary depending on the value of the parent variable (Davison et al., 2016). When the response to the parent item is $X_{kj} = 1$, the information function of the dummy variable Y_{j1} is

$$I(Y_{j1}|\theta) = \frac{(p'_{j1})^2}{p_{j1}(1-p_{j1})}\ . \tag{37}$$

When the response to the parent item is $X_{kj} = 0$, the information function of the dummy variable Y_{j1} is zero, since Y_{j1} is missing. Therefore the expected information function for the dummy variable Y_{j1} is

$$E[I(Y_{j1}|\theta)] = p_{kj}\frac{(p'_{j1})^2}{p_{j1}(1-p_{j1})} + (1-p_{kj}) * 0 = p_{kj}\frac{(p'_{j1})^2}{p_{j1}(1-p_{j1})}\ , \tag{38}$$

where p_{kj} is the probability when $X_{kj} = 1$, and $1 - p_{kj}$ is the probability when $X_{kj} = 0$. When the response to the parent item is $X_{kj} = 0$, the information function of the dummy variable Y_{j0} is

$$I(Y_{j0}|\theta) = \frac{(p'_{j0})^2}{p_{j0}(1-p_{j0})} . \tag{39}$$

When the response to the parent item is $X_{kj} = 1$, the information function of the dummy variable Y_{j0} is zero, since Y_{j0} is missing. Therefore the expected information function for the dummy variable Y_{j0} is

$$E[I(Y_{j0}|\theta)] = p_{kj} * 0 + (1-p_{kj})\frac{(p'_{j0})^2}{p_{j0}(1-p_{j0})} = (1-p_{kj})\frac{(p'_{j0})^2}{p_{j0}(1-p_{j0})} \tag{40}$$

where p_{kj} is the probability when $X_{kj} = 1$, and $1 - p_{kj}$ is the probability when $X_{kj} = 0$. It shows that the expected information function of the child variable is the sum of the expected information functions of the two dummy variables, representing in a formula as

$$E[I(X_j|\theta)] = E[I(Y_{j1}|\theta)] + E[I(Y_{j0}|\theta)] . \tag{41}$$

The Fisher information function does not take into account the situation when the response variable is missing. It is necessary to use the expected information function to evaluate the information of an RCIRT model.

Partchev et al. (2013). Partchev et al. (2013) used the RCIRT approach for the verbal analogies and Raven-like matrices tests. They focused on whether the fast correct responses and the slow correct responses involved different latent intelligence dimensions. Besides response accuracy, response times were recorded and dichotomized into fast and slow categories based on a median split per item. Therefore, the parent variable

represents the speed, $X_{kj} = 1$ if the response is answered in a fast way, and $X_{kj} = 0$ if the response is answered slowly; the child variable represents response accuracy, $X_j = 1$ if the response is correct, and $X_j = 0$ if the response is incorrect. Using the Rasch model, the model for the parent variable is

$$p_{kj} = P\left(X_{kj} = 1 \middle| \theta, b_{kj}\right) = \frac{\exp\left(\theta - b_{kj}\right)}{1 + \exp\left(\theta - b_{kj}\right)}, \tag{42}$$

where θ is the location of the person on the speed dimension, and b_{kj} is the time intensity paramater of the item j. The model for the probability of getting the item correct j is conditional on the speed of response. When the item is answered fast (i.e., $X_{kj} = 1$), the model for the child variable (correct or incorrect) is

$$p_{j1} = P\left(X_j = 1 \middle| X_{kj} = 1, \theta_1, b_{j1}\right) = \frac{\exp\left(\theta_1 - b_{j1}\right)}{1 + \exp\left(\theta_1 - b_{j1}\right)}, \tag{43}$$

where b_{j1} is the difficulty parameter specific to a fast cognitive process of solving the item, and θ_1 is the intelligence under the fast solution process. When the item is answered slow (i.e., $X_{kj} = 0$), the model for the child variable is

$$p_{j0} = P\left(X_j = 1 \middle| X_{kj} = 0, \theta_0, b_{j0}\right) = \frac{\exp\left(\theta_0 - b_{j0}\right)}{1 + \exp\left(\theta_0 - b_{j0}\right)}, \tag{44}$$

where b_{j0} is the difficulty parameter specific to a slow cognitive process of solving the item, and θ_0 is the intelligence under the slow solution process. In Andrich and Kreiner's (2010) models of child variables, the item difficulty parameters are different but the underlying dimension is the same math ability. Whereas in Partchev's (2013) models of child variables, the item difficulty parameters are different, and the underlying

dimensions are different (i.e., θ, θ_1, θ_0), each of which represents a type of intelligence associated with a specific cognitive process. Since every item has a parent variable and a child variable, the likelihood of the response vector for a person can be formulated as (Davison et al., 2016):

$$\prod_j L(X_{kj}, X_j) = \prod_j {p_{kj}}^{x_{kj}}(1-p_{kj})^{1-x_{kj}}[{p_{j1}}^{x_j}(1-p_{j1})^{1-x_j}]^{x_{kj}}[{p_{j0}}^{x_j}(1-p_{j0})^{1-x_j}]^{1-x_{kj}}. \quad (45)$$

This likelihood function can also be formulated in terms of the response vector (X_{kj}, Y_{j1}, Y_{j0}), where Y_{j1}, Y_{j0} are dummy variables, each reflecting a conditional item response model. The information functions of this model are computed the same way as discussed in the example of Andrich and Kreiner (2010).

The models proposed by Andrich and Kreiner (2010) and Partchev et al. (2013) differ in several ways. In Andrich and Kreiner's (2010) model, the parent variable and the child variable correspond to responses on two different items in the test. Both items have the same latent trait, which is the math ability; therefore, this model is unidimensional. In Partchev et al.'s (2013) model, the parent variable and the child variable correspond to two different features, response accuracy and response times, of the same item. This item has three different latent traits, which are response speed (denoted as θ), fast intelligence (denoted as θ_1) and slow intelligence (denoted as θ_0).

In both examples, the response to a child item is dependent on the response to its parent item. Therefore the assumption of local independence is violated. The nature of the dependency could be represented in several, conditional models for the child item. The model for a given person is determined by the person's response to the parent item.

The model can employ different item response functions besides the Rasch model (Partchev & De Boeck, 2012; Partchev et al., 2013).

RCIRT for Reading Comprehension Fluency. To model the reading comprehension fluency through the reading comprehension test, an RCIRT model is proposed. This section discusses the RCIRT used to model reading comprehension fluency proposed in the current study. To measure the reading comprehension fluency using a newly developed reading comprehension test, the response and response times were recorded; therefore, two response variables for each item were generated, among which the parent variable of item j was the response accuracy: $X_{kj} = 1$ if the response was correct and $X_{kj} = 0$ if the response was incorrect. A 2PL IRT model was posited for the parent variable as

$$p_{kj} = P\left(X_{kj} = 1 \middle| \theta, \alpha_{kj}, b_{kj}\right) = \frac{\exp[\alpha_{kj}(\theta - b_{kj})]}{1 + \exp[\alpha_{kj}(\theta - b_{kj})]}, \tag{46}$$

where α_{kj} is the discrimination parameter for item j, b_{kj} is the difficulty parameter for item j, and θ in this model is interpreted as the latent trait of reading comprehension: examinees with high θ estimates are described as good comprehenders and those with low θ estimates are described as poor comprehenders.

The child variable denotes the speed of choosing a correct response, which we call the comprehension efficiency. The child variable of item j, X_j, can be replaced by a dummy variable X_{j1}, where

$$X_{j1} = \begin{cases} X_j, & \text{if } X_{kj} = 1; \\ missing, & \text{if } X_{kj} = 0. \end{cases} \tag{47}$$

Therefore, $X_{j1} = 1$ if a correct response was chosen in a fast way, $X_{j1} = 0$ if a correct response was chosen slowly, and X_{j1} is missing if an incorrect response was chosen. The proposed RCIRT model for the response conditional variable is a 2PL IRT model conditional on $X_{kj} = 1$, with the probability as

$$p_{j1} = P\left(X_{j1} = 1 \middle| X_{kj} = 1, \theta_1, \alpha_{j1}, b_{j1}\right) = \frac{\exp[\alpha_{j1}(\theta_1 - b_{j1})]}{1 + \exp[\alpha_{j1}(\theta_1 - b_{j1})]}, \tag{48}$$

where α_{j1} is the discrimination parameter specific to an efficient process of solving item j, b_{j1} is the difficulty parameter specific to an efficient process of solving item j, and θ_1 is interpreted as the propensity of choosing a correct response fast over choosing a correct response slowly. A high θ_1 estimate indicates a tendency of comprehending efficiently and a low θ_1 estimate indicates a tendency of comprehending inefficiently.

Two information functions can be computed for the dimension θ_1 of the child variable. The first is the information function assuming a value of X_{j1} for each item. This information function for item j can be computed using Equation (37) and a sum of the information functions for all items is the test information. The second is the expected information function taking into account the missing data on the conditional child variable X_{j1}. To compute the expected information function for each θ_1 value, Equation (38) is updated into:

$$E\left[I\left(Y_{j1} \middle| \theta_1\right)\right] = \int_{-\infty}^{\infty} p_{kj}(\theta)\, f(\theta|\theta_1) d\theta * \frac{\left(p'_{j1}\right)^2}{p_{j1}\left(1 - p_{j1}\right)} \tag{49}$$

where $f(\theta|\theta_1)$ is the conditional probability density function for θ given θ_1. The product of $\int_{-\infty}^{\infty} p_{kj}(\theta)\, f(\theta|\theta_1)d\theta$ is equvalent to the proportion of people with a score of θ_1 who can answer the item j correctly. The estimation assumes $(\theta, \theta_1) \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_{\widehat{\theta},\widehat{\theta_1}} \\ \rho_{\widehat{\theta},\widehat{\theta_1}} & 1 \end{pmatrix},$$

and $\rho_{\widehat{\theta},\widehat{\theta_1}}$ is the correlation between estimates from the parent dimension and estimates from the child dimension. The equation (49) assumes that the proportion of people who correctly answer the item will vary as a function of θ_1 because of the correlation between θ and θ_1.

Although sharing some similarities, the RCIRT model and the mixture model are different, since a strict mixture model is conditional on a latent variable (for example, Δ_{ij} in Wang & Xu, 2015), whereas the RCIRT model is conditional on an observed variable.

The RCIRT model has various applications, and the modeling of reading comprehension fluency is one example of its applications. For example, the RCIRT model can be used for computerized adaptive testing, and can be extended to multiple parent variables and multiple children variables if the response model is conditional on a dichotomous function of several parent variables. The RCIRT can also be generalized to polytomous variables. However, it is important to notice that there could be substantial missing data for the RCIRT variables. For example, when a test is very difficult and most examinees only get a small amount of items correctly, the missing data for the RCIRT variables leads to a large conditional standard errors, and as a result the estimates of the

conditional dimension should be interpreted with cautions. Chapter 5 has more discussion about the missingness in the RCIRT.

Chapter III: Method

Sample

A sample of 4,358 students in grade 3-5 from over 50 schools in 13 states participated in the *Multiple-Choice Online Causal Comprehension Assessment* (MOCCA; Carlson, Seipel, & McMaster, 2014) study from February 1st to June 16th, 2016. The sample was recruited through online solicitation of schools. Three different forms were administrated to each grade. A total of 9 forms are used in this study. Each MOCCA form is a 40 item, inferential comprehension test. Table 1.1 shows sample size for each form.

Table 1.1

Sample Size by Grade and Form

Grade	Form	Sample size	Grade total
3	3.1	542	1584
	3.2	520	
	3.3	522	
4	4.1	521	1500
	4.2	500	
	4.3	479	
5	5.1	455	1274
	5.2	416	
	5.3	403	

Table 1.2 shows the percentages of students by grade for each of the demographic variables: gender, ethnicity, free/reduced lunch or not, special education or not, and English-language-learner (ELL) or not.

Table 1.2

Percentages of Students by Grade of Demographic Variables

	Grade	3	4	5
Gender	Female	50.3%	50.3%	47.6%
	Male	49.7%	49.6%	52.1%
Ethnicity	American Indian	1.6%	0.7%	1.3%
	Asian	3.1%	3.0%	1.9%
	Hispanic	8.6%	6.8%	4.8%
	Pacific islander	19.6%	23.4%	23.1%
	White	61.2%	60.3%	62.2%
	Two or more	1.4%	1.7%	2.0%
Free/Reduced Lunch	Yes	38.7%	33.9%	29.8%
	No	33.0%	30.5%	35.4%
Special Ed	Yes	10.2%	11.4%	12.7%
	No	86.0%	84.6%	82.4%
ELL	Yes	9.6%	10.9%	6.3%
	No	76.3%	75.9%	82.4%

Instrument

MOCCA is a multiple-choice, online, causal comprehension assessment, which was designed to identify comprehension processes used during reading of narrative texts for 3rd grade, 4th grade and 5th grade (Carlson et al., 2014). Every MOCCA test consists

of 40 short stories, each of which contains a title and seven sentences with the sixth sentence deleted/missing. The consistent format of items is beneficial for the current study. There are 9 different online forms, with 3 different online forms per grade: form 3.1, 3.2, 3.3 for grade 3, form 4.1, 4.2, 4.3 for grade 4, and form 5.1, 5.2, 5.3 for grade 5. In addition, each form has one version of forward order of items and another version of backward order of items. Each response type is randomized per item. Participants are randomly assigned to one of six versions (3 forms x 2 item orders) of the test at their grade.

Participants are required to choose one of three alternative response types to fill in the deleted sentence. The correct answer represents a causally coherent inference, which closes the causal gap between the 5th and 7th sentences. The incorrect options represent each of the two informative distractors: *Paraphrase* and *Lateral Connection*. The Paraphrase distractor repeats the main character's goal or a combination of the goal and subgoal statements presented in the text. The Lateral Connection distractor elaborates on the information in the text.

Procedure

MOCCA was administered through an online platform. Students took the MOCCA test with computers or tablets in their classrooms or in the school computer labs, under the supervision of a trained project staff member. Before the test, students were given instructions that they were going to read several short stories, that each story has a missing sentence, and that their job was to pick one out of three sentences below each story that best completed the story. They could click on the answer choices to see

the sentences within the context of the story. There was a sample item to illustrate how to complete tests and navigate through the program. Students were not able to skip items. There was no time limit of completing the test. After completing all items, students could review their previous responses if time allowed. Students with less than ten responses were eliminated from the sample.

Analysis

The IRT models proposed by the current study follow two approaches respectively. The first approach uses the polytomous response modeling. Using the polytomous scoring, responses to each item are scored into three categories: 0 = incorrect response, 1 = slow and correct response, 2 = fast and correct response. As I discussed, fast while being correct is defined as fluent/automatic in the reading literature. The polytomous models include the nominal response model (NRM), partial credit model (PCM), generalized partial credit model (GPCM) and graded response model (GRM). The first model treats the response categories as nominal and the other models treat the response categories as ordinal. The reason I include both the nominal model and the ordered models is that there might be uncertainty of the cognitive processes behind the fast responses and the slow responses. If the fast-correct responses involve more/higher ability than the slow-correct responses, a model with ordinal categories could be appropriate. If the abilities underlying the fast-correct and slow-correct responses differ in nature, a model with nominal categories could be appropriate. The parameterization of the nominal model helps to identify the empirical ordering of the categories by inspecting the values of item discrimination parameter (de Ayala, 2009). The global fit of the

various polytomous models will be compared. A unidimensional 2PL IRT model of the parent variables (accuracy data) will be used as the baseline model to compare against, in terms of information functions and other criteria. The theta estimates of polytomous models will be compared and correlated with the theta estimates of the 2PL IRT model.

The second approach uses the RCIRT model and scores each of the original responses into two dichotomous response variables. The former is the parent variable, which indicates whether the response is correct or incorrect (0 = incorrect response, 1 = correct response), while the latter indicates whether the correct response is fast or slow. The current study refers to the second response variable as the child variable. If the parent variable corresponds to a correct response that is chosen with a fast process, the child variable is coded as 1; if the parent variable corresponds to a correct response that is chosen in a slow process, the child variable is coded as 0. For those whose answer the items incorrectly, their responses will be missing in the child variables. The person location estimates of the parent variables assess the tendency of being correct in the dimension of comprehension ability. The person location estimates of the child variables assess the propensity of the examinee to be a fast respondent or slow when solving the item correctly. The underlying latent trait of the child variable is interpreted as the comprehension efficiency, a dimension reflecting the rate at which one can arrive at a correct response.

Before proposing the RCIRT models, I want to make sure the missingness in the child variable is not problematic by checking the sum of accurate scores across forms. For the total sample, 92.1% students answered more than 10 items correctly out of 40

items in a form; 62.7% students answered more than 20 items correctly out of 40 items in a form. Table 2 shows the percentages of students who answered more than 10, 15 and 20 items correctly for each form. We can see that more than half of the students answered at least 20 items correctly for each form. Therefore, the missingness in the child variables is not a big concern in the RCIRT model. Chapter 5 has more discussion about the missingness.

Table 2

Percentages of Students with More Than 10, 15, and 20 Correct Item Responses for Different Forms

Form	> 20 correct	> 15 correct	> 10 correct
3.1	0.52	0.67	0.87
3.2	0.50	0.72	0.88
3.3	0.55	0.74	0.89
4.1	0.66	0.81	0.93
4.2	0.66	0.83	0.93
4.3	0.65	0.77	0.93
5.1	0.73	0.87	0.96
5.2	0.74	0.85	0.96
5.3	0.70	0.82	0.96

For the conditional models, the next chapter will compare the results for a unidimensional 2PL IRT model with only parent variables, a unidimensional 2PL IRT model with only child variables, a unidimensional 2PL IRT model with combined parent variables and child variables (which is referred to as an automaticity model), and a two-

dimensional IRT model with comprehension ability being the first dimension and comprehension efficiency the second dimension.

The same analyses will be applied to each of the three forms at each grade. If the observed pattern is consistent within grade, measures could be averaged within grade. If there is significant difference across forms, I will then compare forms within grade, since the forms within grade were designed to be parallel; and I will compare forms across grade, since it could be a result of grade differences.

Measures

This section discusses the evaluation criteria used for both approaches as dependent variables. I follow a similar procedure and analyze each of the nine test forms thoroughly using the measures discussed below:

Polytomous Models. The measures that I will use to compare the polytomous models and the unidimensional 2PL IRT model with only accuracy variables are:

1) Global model fit statistics, including -2 log likelihood (-2LL), RMSEA, and information criteria AIC (Akaike, 1974), BIC (Schwarz, 1978).
2) Marginal reliability. The marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984) uses the classical definition of reliability as proportion of variance in the test score due to true score:

$$\rho = \frac{\sigma_\theta^2 - \bar{\sigma}_e^2}{\sigma_\theta^2} = 1 - \frac{\bar{\sigma}_e^2}{\sigma_\theta^2}\ , \tag{50}$$

where the true score variance is computed as the test score variance minus error variance. There are two ways to compute the marginal reliability coefficient. The

theoretical reliability considers the theoretical distribution of the latent trait as the standard normal distribution, therefore $\sigma_\theta^2 = 1$ and Equation (50) becomes

$$\rho = 1 - \overline{SE}_\theta^2, \tag{51}$$

where the error variance is approximated by the average squared standard error $\overline{SE}_\theta^2$. The empirical reliability considers the standard errors for the estimated sample scores, and is computed in the following formula:

$$\rho = 1 - \frac{\overline{SE}_{\hat{\theta}}^2}{\hat{\sigma}_\theta^2} \tag{52}$$

where $\hat{\sigma}_\theta^2$ is the variance of estimated theta scores and $\overline{SE}_{\hat{\theta}}^2$ is the average standard error variance of the estimated theta scores.

3) Information function. The information function of the polytomous models will be compared with the information functions of a unidimensional 2PL IRT model of the parent variables only. I will compare the average information function measured over the theta continuum from -2.8 to 2.8. Graphs and tables comparing information functions across the theta continuum from -2.8 to 2.8 will be shown.
4) Correlations of theta values. The scatter plots of theta values of each of the polytmous models against the unidimensional 2PL model will be plotted to check the degree of overlapping.
5) Expected and empirical category characteristic plots. The expected and empirical frequency of people choosing each of the categories will be plotted against their theta values respectively. It shows how well the estimates of the selected model represent the data.

Conditional Models. The measures that I will use to compare the 2PL IRT model with only parent variables, the 2PL IRT model with only child variables, the 2PL IRT model with combined parent variables and child variables, and the MIRT model with combined parent variable and child variables are

1) Global model fit statistics: including -2LL, RMSEA, AIC, BIC.
2) Marginal and subscore reliability. For the unidimensional 2PL IRT model, theoretical and empirical reliability coefficients are used to compute the marginal reliability. For the MIRT model, the subscore reliability is computed (Thissen & Wainer, 2001).
3) Information function. At each grade I compute and compare the information function of parent variables, information function of child variables, and expected information function of child variables. Two information functions can be computed for the child variables: the Fisher information function for a child variable is computed using Equation (37), while the expected information function for a child variable is computed using Equation (49). The expected information function taking into account missingness in the child dimension would provide useful and practical information concept in that application. The test information for the conditional dimension, given all the responses are present, is a rather rare situation in which the examinee answered every item correctly. Also the information functions of the combined parent and child variables are computed in the unidimensional case and in the multidimensional case. Specifically, in the MIRT model the information function is computed in the direction of latent trait. Plots are presented to compare the Fisher information or expected information curves in the unidimensional case.

4) Average conditional standard error for each person over the theta values of (-2.5, -1.5, -0.5, 0.5, 1.5, 2.5).
5) Correlation of the theta estimates of ability dimension and efficiency dimension will also be computed to examine the extent to which the two latent traits are overlapped.

Connecting Polytomous Models and Conditional Models. To examine the relationship of latent traits underlying the polytomous model and the RCIRT model, correlation of the theta estimates of the polytomous model and those of the RCIRT model will be obtained.

Chapter IV: Results

This chapter presents the results for the polytomous models and the conditional models. Before presenting the results, the assumption of response times is checked against the presence of abnormal behaviors in the data, such as rapid guessing, that could bias the results. In the end, a summary is presented to compare the chosen polytomous model and the chosen RCIRT model.

Response Times

It is assumed that the examinee will give adequate effort when taking the MOCCA test. I would like to check the response times to see whether the assumption is violated. For each form, Figure 1 and Figure 2 present the histograms of log-transformed response times for the item-person combinations for all the item responses and for all the correct item responses respectively. According to previous studies (Meyer, 2010; Schnipke & Scrams, 1997; Wang & Xu, 2015), response times obtained through the mixture of solution behavior and rapid guessing would display a bimodal distribution. In Figure 1 and 2, there is no pattern of bimodal distribution detected in the response time distribution for all the item responses or for all the correct item responses in each form.

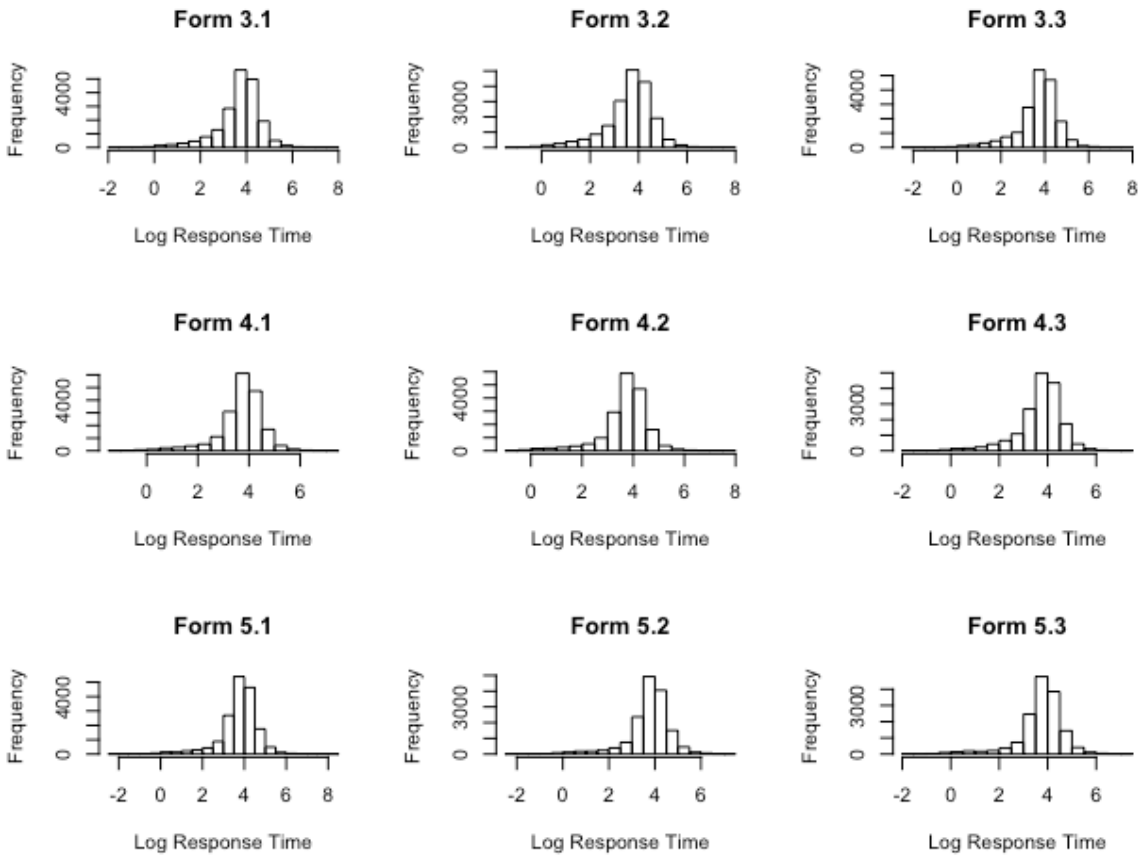

Figure 1. Histogram of log-transformed response times for all item-person combinations in each form

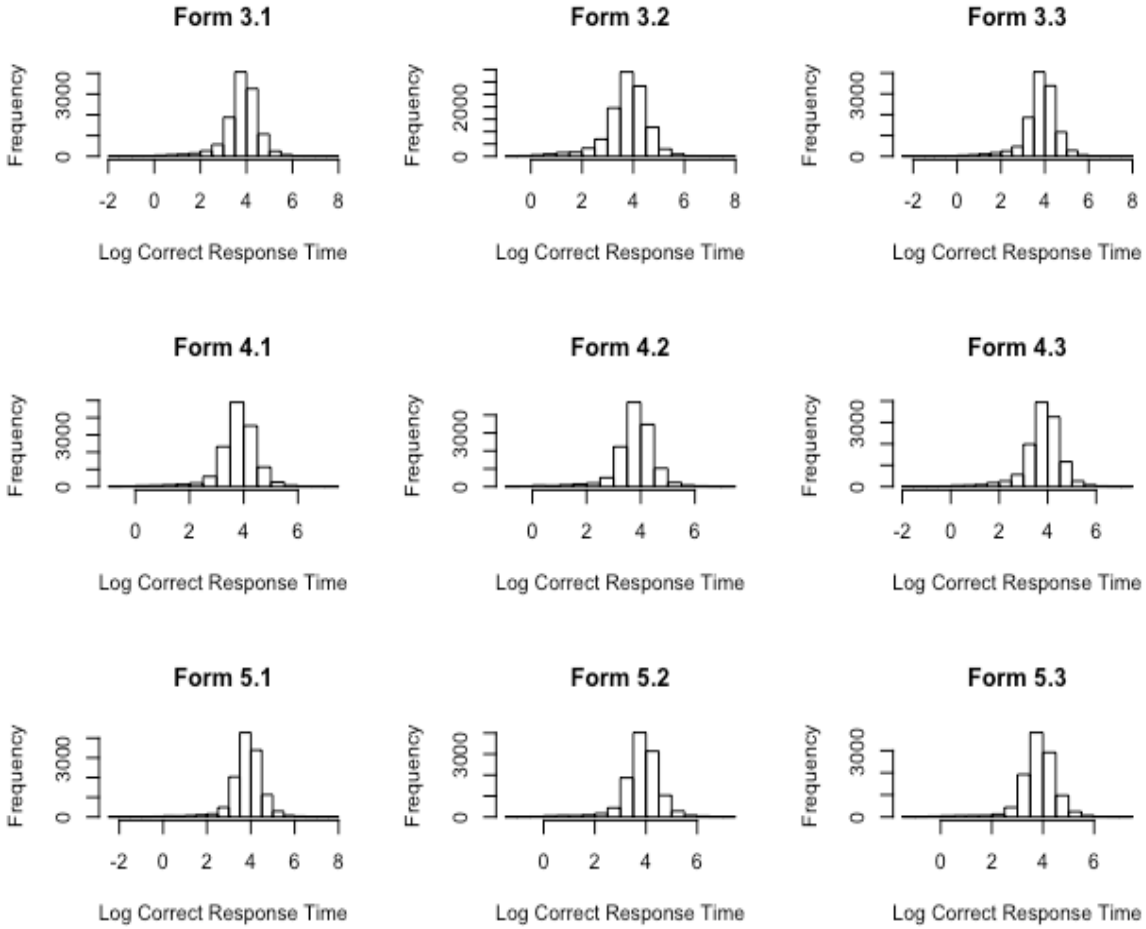

Figure 2. Histogram of log-transformed response times for all correct-item-person combinations in each form

The person-level response time trajectories are plotted to further identify examinees' testing behaviors. The person-level response time trajectories of all forms display a similar pattern; therefore I only use Form 3.1 as an example in Figure 3 and Figure 4. Figure 3 shows the accumulated number of items answered as a function of the accumulated log-transformed response time for examinees of Form 3.1. Each line indicates the trajectory of items vs. response time for one examinee. If an examinee used

a rapid guessing behavior starting at a certain point, his/her trajectory would turn upward sharply, indicating an extremely short time interval of solving a set of items. Figure 3 shows the average trajectory is approximately linear, suggesting that examinees of Form 3.1 worked at a nearly constant speed throughout the test. In other words, the examinees tend to use one type of behaviors during the test, rather than switching to different testing behaviors. The "slopes" of the trajectory suggest that most examinees took the solution behavior throughout the test.

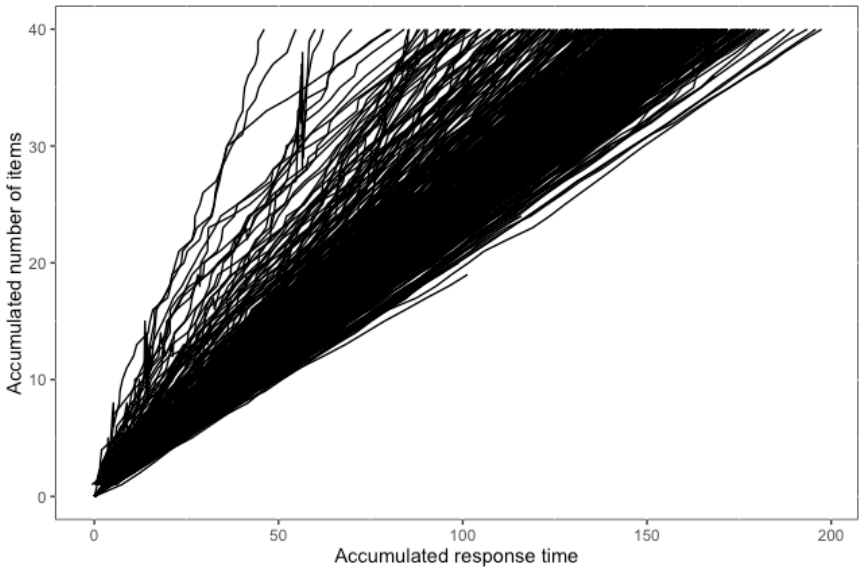

Figure 3. Accumulated number of items answered against the accumulated log-transformed response time for each examinee of Form 3.1.

Similarly, Figure 4 presents the accumulated number of items answered correctly as a function of the accumulated log-transformed response time of corresponding correct item responses for examinees of Form 3.1. No sharp upward trajectories are detected, suggesting that it is unlikely that correct answers are results of rapid guessing behaviors.

After examining the distributions and trajectories of response times in Figure 1 – 4, examinees with abnormal behaviors, such as rapid guessers, are not a major concern for our data. No evidence of violation of the assumption is found.

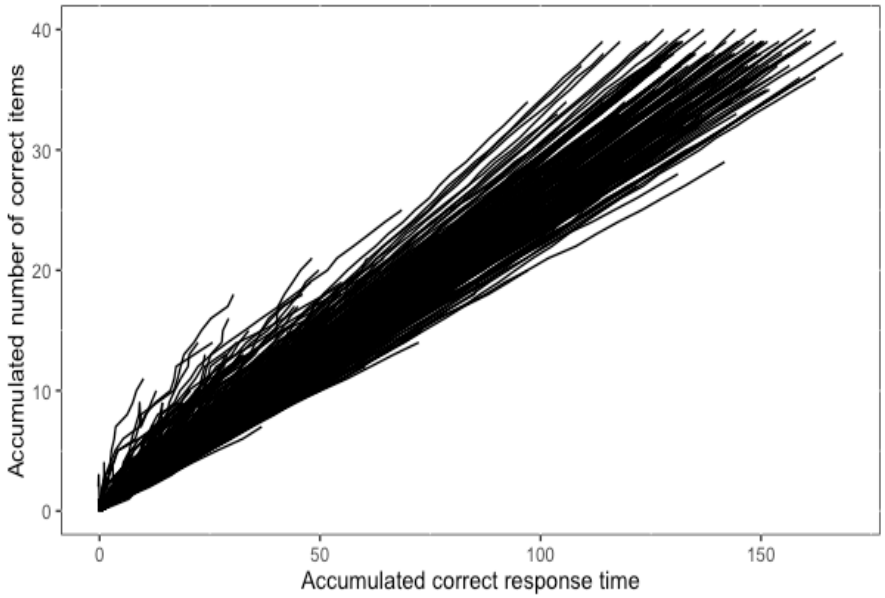

Figure 4. Accumulated number of items answered correctly against the accumulated log-transformed response time of the correct item responses for each examinee of Form 3.1.

Polytomous Models

The ANOVA tests were conducted to check if there is a difference of mean theta estimates between forms within grade. The ANOVA tests for within grade show that for each of the models, there is no significant difference of theta estimates between different forms within each grade.

The initial analyses indicate the nominal response model doesn't perform well on the measuring criteria. For example, for the fit indices, the nominal response models have the highest RMSEA. The item discrimination parameters of the nominal response models

are ordered increasingly (i.e., $a_0 < a_1 < a_2$) for items across forms, which are expected when items are truly ordinal (de Ayala, 2009). Also, correct and incorrect are clearly ordered categories, so a purely nominal model is not appropriate. Therefore, I only consider polytomous models for ordinal responses, including the partial credit model, generalized partial credit model, and graded response model. The decision of choosing the ordinal response models over the nominal response model is in agreement with the theory of reading comprehension fluency.

The theoretical marginal reliabilities and empirical marginal reliabilities of different models for each form are shown in Table 3 and Table 4. The differences of the two reliabilities were discussed in Chapter 3. The unidimensional 2PL model has lower marginal reliabilities (either theoretical or empirical) than any of the polytomous models. Among them, the graded response models maintain the highest marginal reliabilities across forms.

Table 3

Theoretical Reliabilities of PCM, GPCM, GRM, and 2PL Models across Forms

Form	PCM	GPCM	GRM	2PL
3.1	0.931	0.936	0.942	0.912
3.2	0.919	0.927	0.933	0.907
3.3	0.925	0.931	0.938	0.902
4.1	0.935	0.938	0.944	0.892
4.2	0.937	0.940	0.945	0.893
4.3	0.931	0.935	0.940	0.894
5.1	0.930	0.936	0.942	0.887
5.2	0.937	0.941	0.946	0.874
5.3	0.940	0.944	0.950	0.883

Table 4

Empirical Reliabilities of PCM, GPCM, GRM, and 2PL Models across Forms

Form	PCM	GPCM	GRM	2PL
3.1	0.919	0.925	0.932	0.904
3.2	0.905	0.915	0.922	0.897
3.3	0.909	0.916	0.925	0.891
4.1	0.925	0.928	0.934	0.887
4.2	0.927	0.930	0.935	0.887
4.3	0.917	0.922	0.928	0.886
5.1	0.925	0.931	0.936	0.887
5.2	0.929	0.935	0.938	0.874
5.3	0.934	0.939	0.945	0.885

I averaged the theoretical marginal reliabilities and the empirical mariginal reliabilities over forms within each grade to examine the changes of reliability across grade levels. The results in Table 5 suggest that as grade goes up, the reliability of each of the polytomous models increases whereas the reliability of the 2PL model decreases. The upward trend of reliability of the polytomous models is expected since there should be more responses that are accurate (as shown in Table 2) and fast as the grade goes up. This suggests the advantage of a polytomous model over the 2PL model since it is able to capture the information reflected by changes of grade.

Table 5

Marginal Reliabilities of PCM, GPCM, GRM, and 2PL Models across Grades

	Grade	PCM	GPCM	GRM	2PL
Theoretical	3	0.925	0.931	0.938	0.907
	4	0.934	0.937	0.943	0.893
	5	0.936	0.940	0.946	0.881
Empirical	3	0.911	0.919	0.926	0.897
	4	0.923	0.927	0.932	0.887
	5	0.929	0.935	0.940	0.882

To check if the different models measure the same latent trait, I checked the Pearson correlations of the theta estimates of different polytomous models and the unidimensional 2PL model. The correlation matrices of theta estimates for each form are displayed in Table 6.1. The theta estimates of different polytomous models are all highly correlated. The correlations of theta estimates of the unidimensional 2PL model with those of polytomous models are above .80. This suggests the abilities measured in the polytomous models are not exactly the same as the ability measured in the unidimensional 2PL model, but they are highly similar. The correlation between the unidimensional 2PL model and each of the polytomous models decreases as the grade goes up. For forms within each grade, the correlations between different polytomous models are very close, whereas the correlations of each polytomous model with the unidimensional 2PL model vary across forms. After capturing the response time information, the estimation of theta tends to be more reliable across forms within grade.

Table 6.1

Correlation of Thetas for PCM, GPCM, GRM, and 2PL Models of Different Forms

Form		PCM	GPCM	GRM	2PL
3.1	PCM	1			
	GPCM	0.997	1		
	GRM	0.991	0.996	1	
	2PL	0.875	0.872	0.879	1
3.2	PCM	1			
	GPCM	0.996	1		
	GRM	0.991	0.996	1	
	2PL	0.854	0.849	0.864	1
3.3	PCM	1			
	GPCM	0.997	1		
	GRM	0.991	0.996	1	
	2PL	0.827	0.822	0.830	1
4.1	PCM	1			
	GPCM	0.998	1		
	GRM	0.995	0.997	1	
	2PL	0.836	0.832	0.838	1
4.2	PCM	1			
	GPCM	0.998	1		
	GRM	0.994	0.997	1	
	2PL	0.854	0.851	0.856	1
4.3	PCM	1			
	GPCM	0.998	1		
	GRM	0.993	0.996	1	
	2PL	0.835	0.829	0.834	1
5.1	PCM	1			
	GPCM	0.997	1		
	GRM	0.992	0.997	1	
	2PL	0.828	0.817	0.819	1
5.2	PCM	1			
	GPCM	0.996	1		
	GRM	0.993	0.997	1	
	2PL	0.823	0.810	0.810	1
5.3	PCM	1			
	GPCM	0.996	1		
	GRM	0.992	0.996	1	
	2PL	0.829	0.815	0.820	1

The correlation between theta estimates of polytomous models and the dichotomous 2PL model serves as the evidence of validity, I further checked whether difference in validity can be accounted for by the marginal reliability. Table 6.2 below shows the correlations corrected for unreliability as evidence of validity. After corrections for unreliability the correlations increase, especially for grade 3. The theta estimates of polytomous models and the dichotomous 2PL model are not perfectly correlated, which suggests that the latent traits measured in the polytomous models are not exactly the same as the latent trait measured in the unidimensional 2PL model, but they are highly similar.

Table 6.2

Corrected Correlation of Thetas for PCM, GPCM, GRM Against 2PL Models

Form		PCM	GPCM	GRM
3.1	2PL	0.950	0.944	0.948
3.2	2PL	0.935	0.926	0.939
3.3	2PL	0.905	0.897	0.902
4.1	2PL	0.915	0.910	0.913
4.2	2PL	0.934	0.929	0.932
4.3	2PL	0.915	0.907	0.910
5.1	2PL	0.912	0.897	0.896
5.2	2PL	0.909	0.893	0.891
5.3	2PL	0.910	0.893	0.895

Figures 5.1-5.9 display the scatter plots of theta values for each polytomous model against the unidimensional 2PL model for all the forms. The scatter plots provide further information about where the disparity of the theta estimates is. The theta values are scattered on the high end of the theta scale. On the low end of the theta scale, the theta estimates of a polytomous model and a 2PL model are highly overlapped. This

suggests that the accuracy and speed work together for examinees of lower ability but not the same for examinees of higher ability. Unbiased external criteria, which measure more than accuracy, are needed to determine if students with the same accuracy abilities are actually different. Teacher evaluation could be one example of the external criterion.

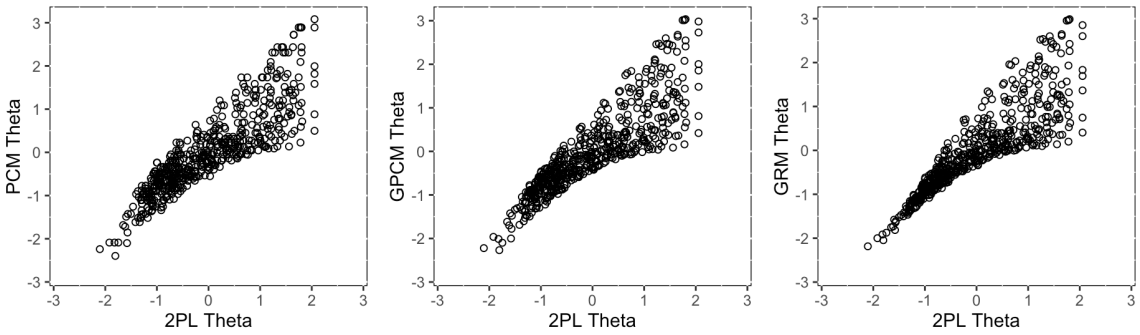

Figure 5.1. Scatter plots of theta values for the polytomous models against the 2PL model of Form 3.1

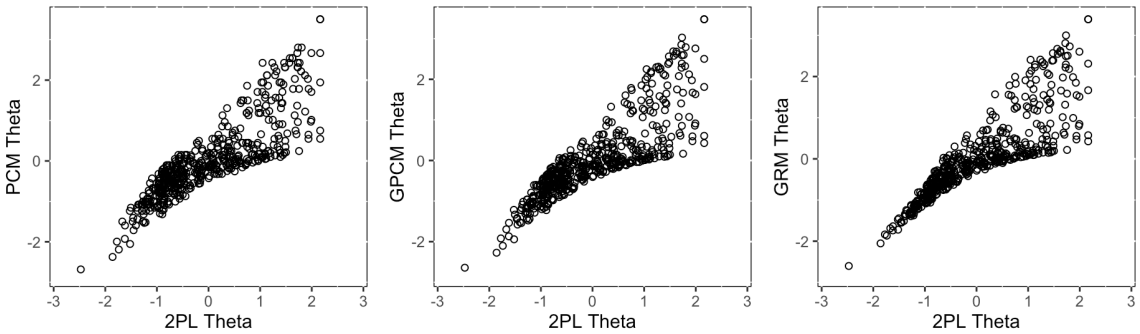

Figure 5.2. Scatter plots of theta values for the polytomous models against the 2PL model of Form 3.2

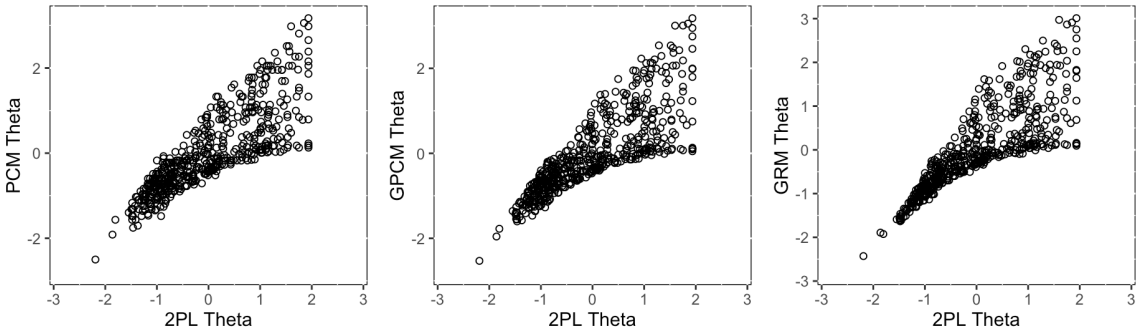

Figure 5.3. Scatter plots of theta values for the polytomous models against the 2PL model of Form 3.3

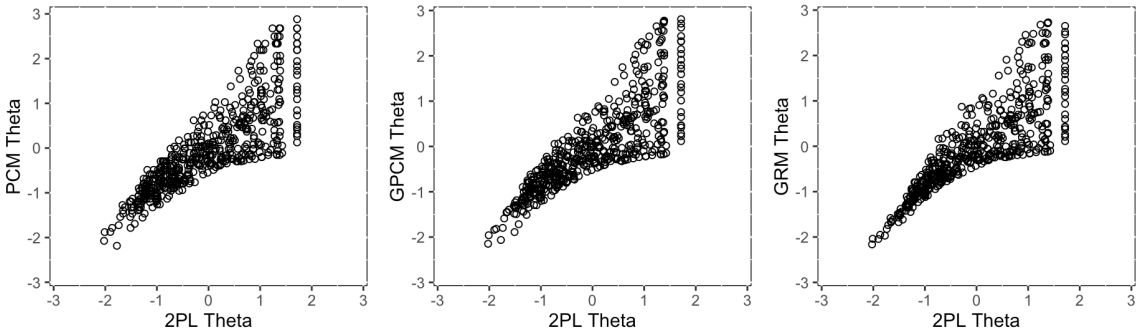

Figure 5.4. Scatter plots of theta values for the polytomous models against the 2PL model of Form 4.1

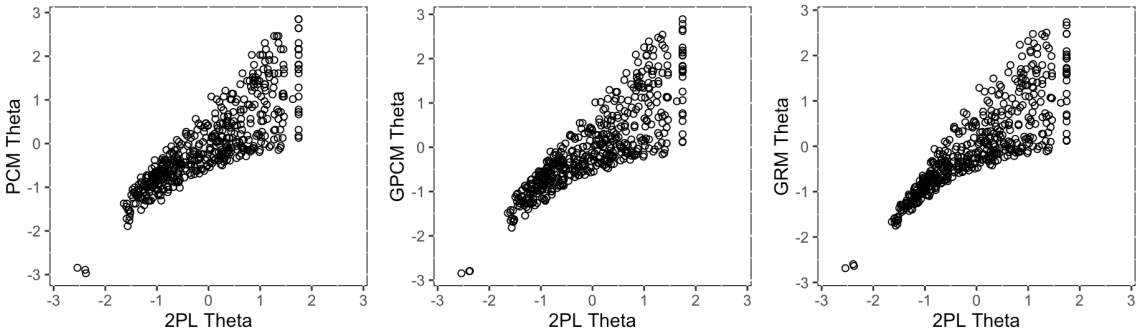

Figure 5.5. Scatter plots of theta values for the polytomous models against the 2PL model of Form 4.2

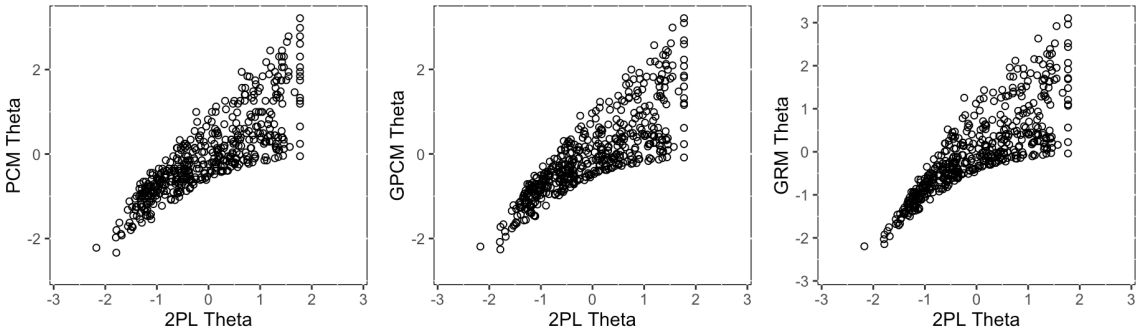

Figure 5.6. Scatter plots of theta values for the polytomous models against the 2PL model of Form 4.3

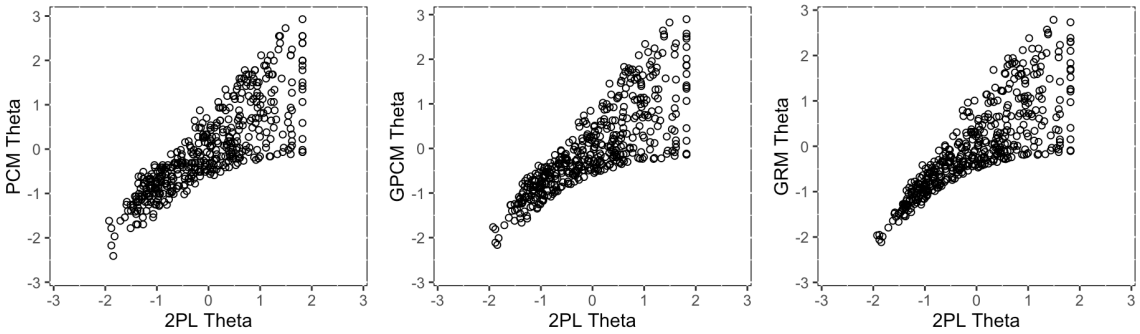

Figure 5.7. Scatter plots of theta values for the polytomous models against the 2PL model of Form 5.1

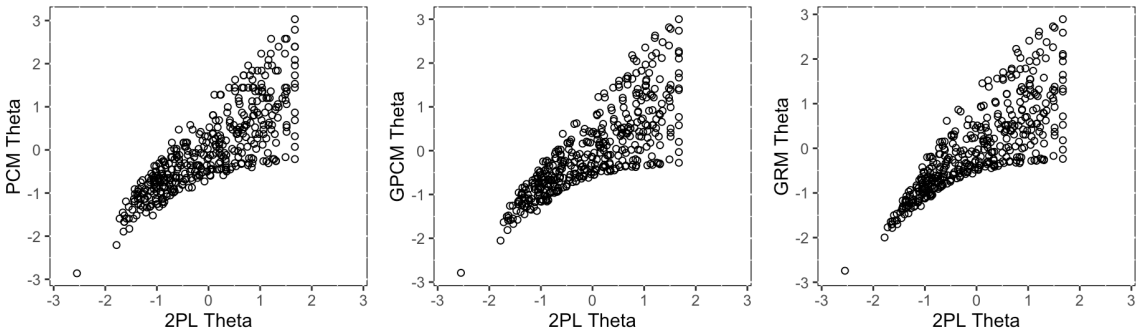

Figure 5.8. Scatter plots of theta values for the polytomous models against the 2PL model of Form 5.2

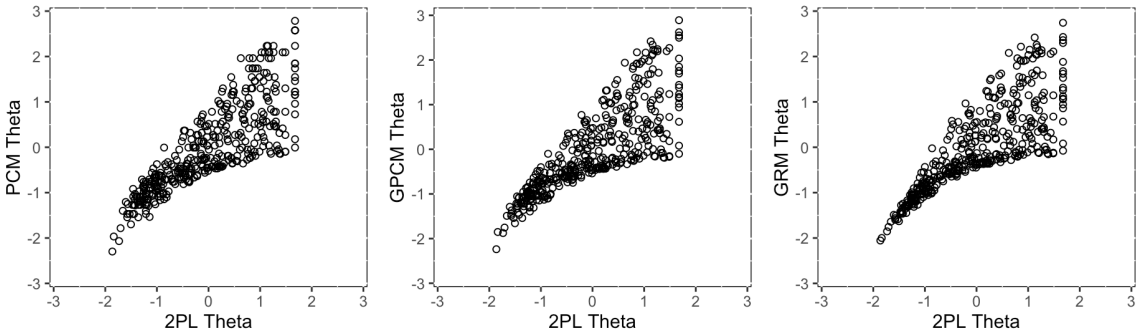

Figure 5.9. Scatter plots of theta values for the polytomous models against the 2PL model of Form 5.3

Table 7 shows the model fit indices of each form, including AIC, BIC, -2 Log likelihood (-2LL) and RMSEA. Among all the polytomous models, the graded response model consistently fits the best for all the forms. The graded response models have the lowest AIC, BIC and -2LL statistics.

Table 7

AIC, BIC, -2LL and RMSE of Polytomous Models of Different Forms

Form		AIC	BIC	-2LL	RMSEA
3.1	PCM	37068.90	37416.81	36906.90	0.02
	GPCM	36872.46	37387.89	36632.46	0.02
	GRM	36545.86	37061.29	36305.86	0.02
3.2	PCM	36752.58	37097.14	36590.58	0.02
	GPCM	36571.18	37081.64	36331.18	0.02
	GRM	36273.07	36783.53	36033.06	0.02
3.3	PCM	35851.74	36196.61	35689.74	0.01
	GPCM	35685.66	36196.58	35445.66	0.02
	GRM	35378.48	35889.40	35138.48	0.02
4.1	PCM	36522.39	36867.11	36360.40	0.02
	GPCM	36450.93	36961.62	36210.92	0.02
	GRM	36176.50	36687.19	35936.50	0.02
4.2	PCM	34852.35	35193.73	34690.34	0.02

	GPCM	34745.25	35251.00	34505.24	0.02
	GRM	34492.21	34997.96	34252.20	0.02
4.3	PCM	33321.22	33659.13	33159.22	0.02
	GPCM	33224.70	33725.30	32984.70	0.02
	GRM	33000.73	33501.34	32760.74	0.02
5.1	PCM	33397.22	33730.97	33235.22	0.02
	GPCM	33190.07	33684.51	32950.08	0.02
	GRM	32951.48	33445.92	32711.48	0.02
5.2	PCM	29181.37	29507.86	14509.69	0.03
	GPCM	28994.66	29478.34	28754.66	0.03
	GRM	28828.90	29312.59	28588.90	0.03
5.3	PCM	28292.27	28616.18	28130.26	0.02
	GPCM	28088.54	28568.41	27848.54	0.02
	GRM	27815.18	28295.05	27575.18	0.02

The average test information function along the theta scale of [-2.8, 2.8] was calculated for each form and the results are shown in Table 8. The graded response models have the highest average test information functions across forms, whereas the unidimensional 2PL models have the lowest average information functions across forms.

Table 8

Average Information of PCM, GPCM, GRM, and 2PL Model

Form	PCM	GPCM	GRM	2PL
3.1	10.77	11.53	12.97	10.51
3.2	9.21	10.00	11.24	9.23
3.3	9.87	10.60	12.10	10.43
4.1	11.41	11.95	13.42	10.81
4.2	11.70	12.29	13.63	10.67
4.3	10.67	11.25	12.55	10.92
5.1	10.60	11.60	12.95	10.18
5.2	11.76	12.90	14.10	11.07
5.3	12.34	13.48	15.14	11.56

Figures 6.1-6.9 show the test information functions of different models for each form. Note that the dimension of the unidimensional 2PL model and the dimension of the polytomous models are not identical. Two facts are observed consistently across forms: first, the information function of the graded response model is high over a broader range of the theta scale. The information function provided by the unidimensional 2PL model is concentrated on the lower end of the scale; however, on the higher end of the theta scale, the 2PL model doesn't provide as much information as each of the polytomous models does. In other words, the 2PL model provides minimum information for students with high abilities, thus using this model to estimate fluent students would not yield precise estimates and the corresponding standard errors would be comparatively large. To estimate students with theta above zero, the graded response model provides the maximum information compared with other models. Second, the information function of the graded response model centers at zero, whereas the information function of the unidimensional 2PL model centers around -1. In practice, the majority of the population is around the theta value of zero, thus a model providing more information at this interval is more valuable than models providing more information at other theta intervals. Therefore, we prefer the graded response model since it provides the highest information function and it outperforms other models when estimating students with theta above zero.

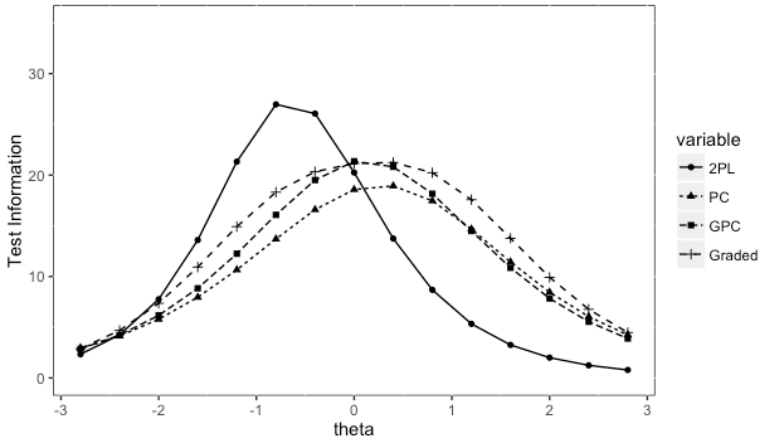

Figure 6.1. Information curve of models of Form 3.1

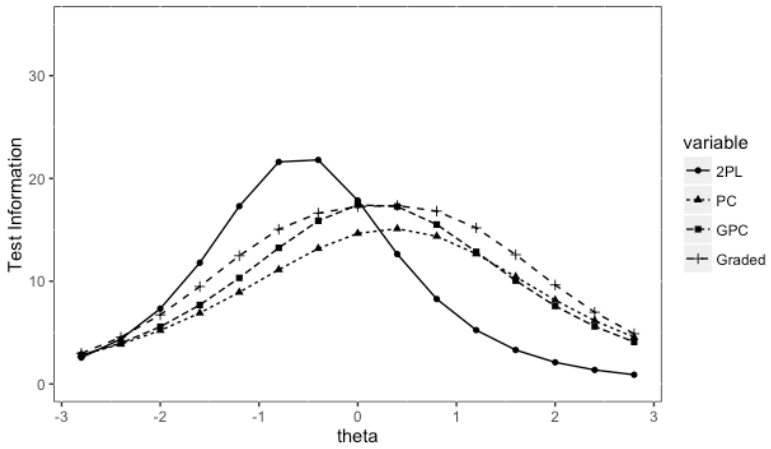

Figure 6.2. Information curve of models of Form 3.2

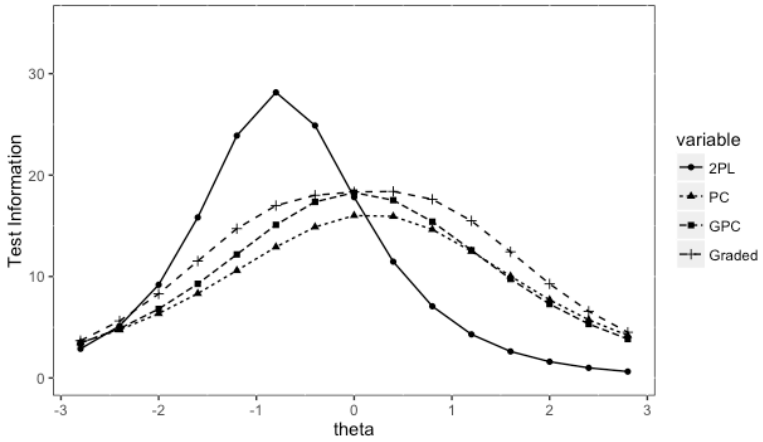

Figure 6.3. Information curve of models of Form 3.3

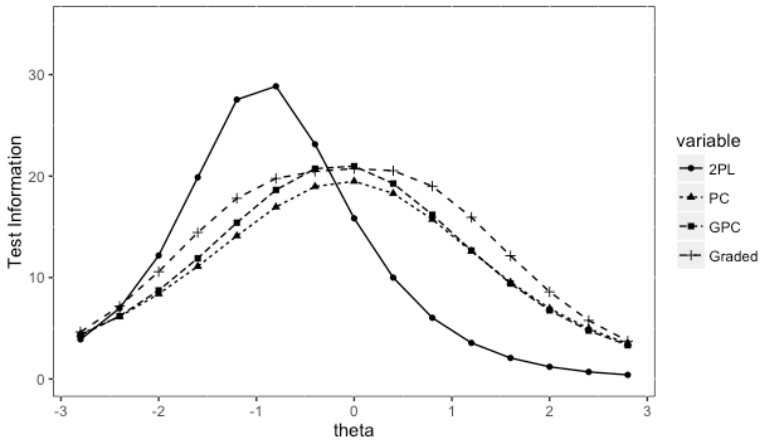

Figure 6.4. Information curve of models of Form 4.1

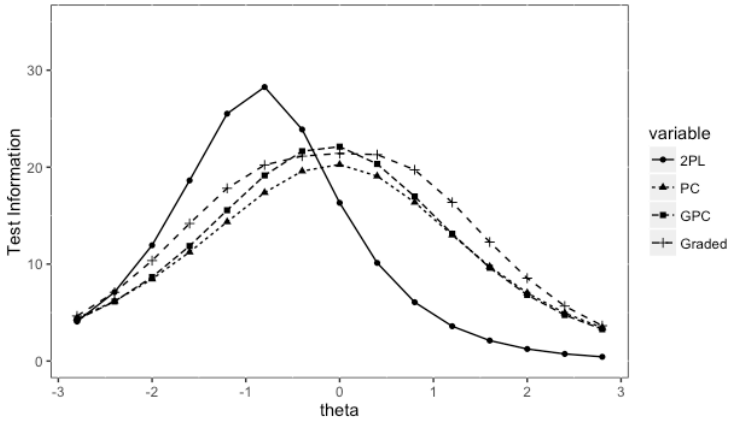

Figure 6.5. Information curve of models of Form 4.2

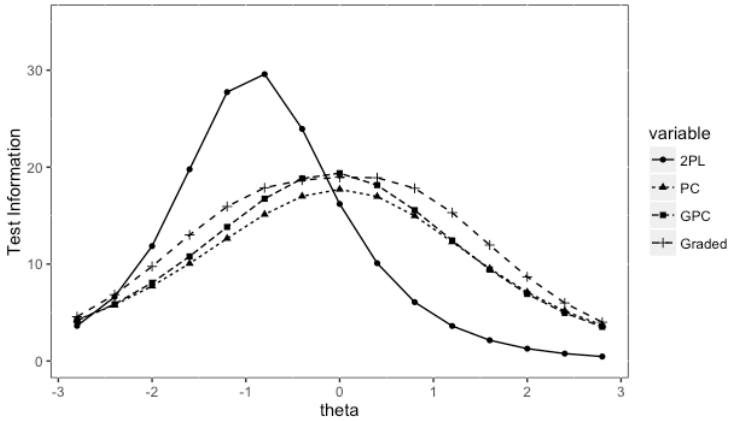

Figure 6.6. Information curve of models of Form 4.3

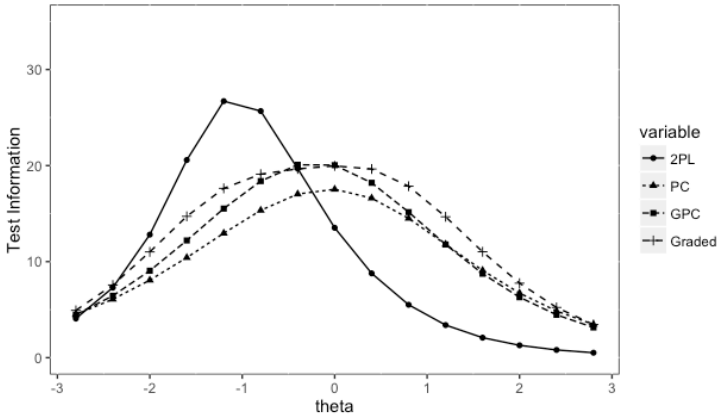

Figure 6.7. Information curve of models of Form 5.1

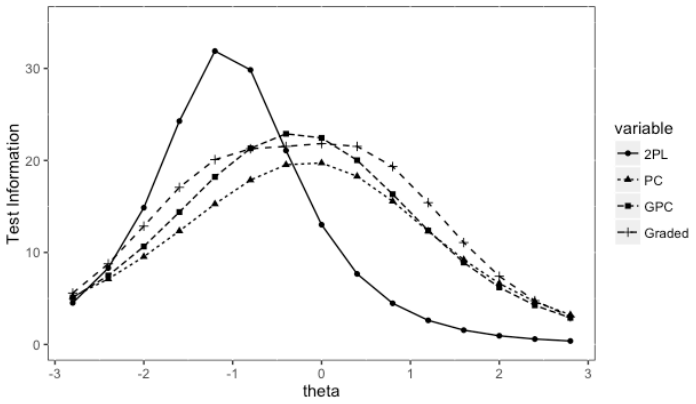

Figure 6.8. Information curve of models of Form 5.2

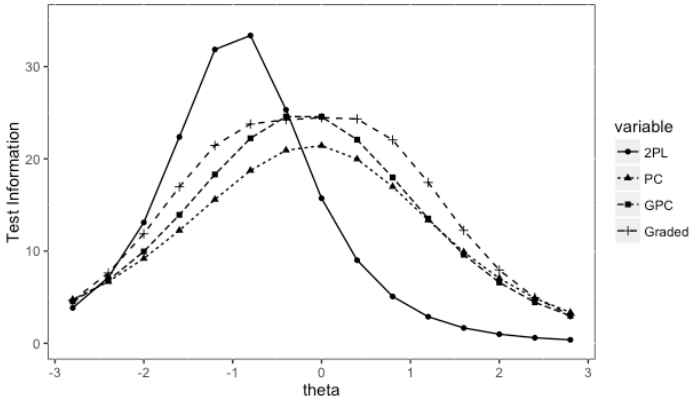

Figure 6.9. Information curve of models of Form 5.3

To explain the center and shape of the test information function of the graded response model, I further checked its category information functions at the test level for the form average of each grade. Figures 7.1-7.3 below show that the information functions for the incorrect response and the fast-correct response are unimodal, while the information function for the slow-correct response is bimodal with a minimum around theta of 0. Because of the shape and distribution of the category information functions, the test information of the graded response model is symmetric at the vertical line of $\theta = 0$, and is high over a broader range of θ.

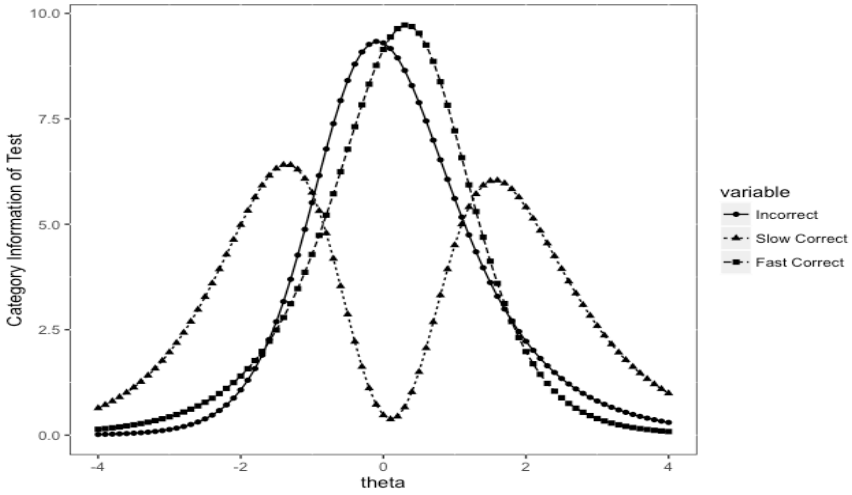

Figure 7.1. Category information curves of Grade 3 forms averaged for the graded response model

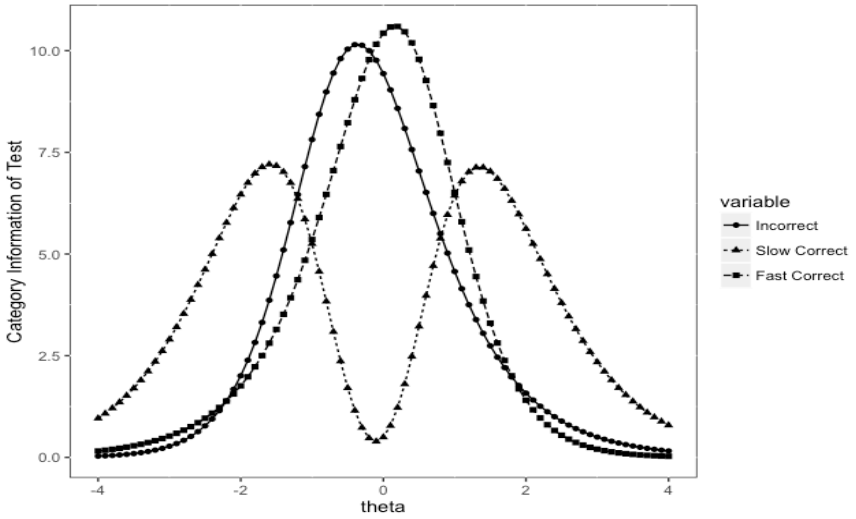

Figure 7.2. Category information curves of Grade 4 forms averaged for the graded response model

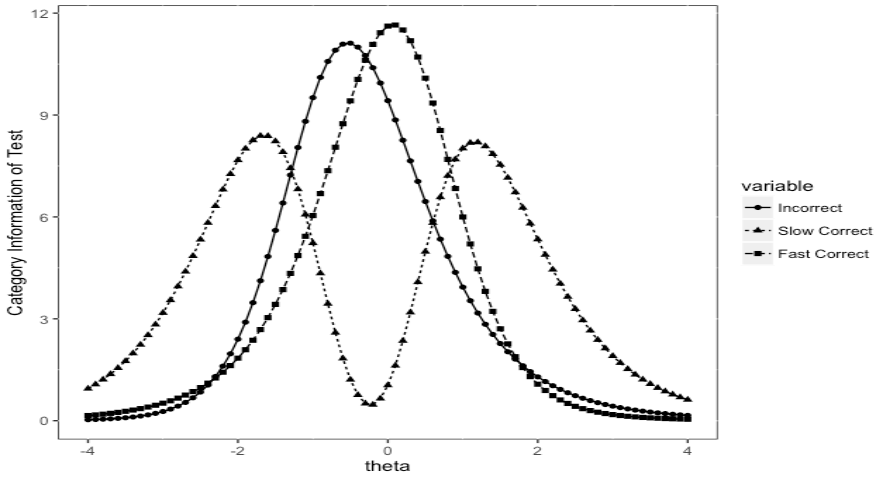

Figure 7.3. Category information curves of Grade 5 forms averaged for the graded response model

The graded response model has the best model fit, the highest reliabilities, and the highest average information functions across the theta scale, thus it seems to be a proper model to be used in the polytomous cases. Using the graded response model, I checked

the expected and empirical category functions separately for each form. Figures 8.1 – 8.6 represent the expected and the empirical category characteristic plots for each grade respectively. The vertical axis of the expected category characteristic curves reflects the expected frequency of being in each of the categories for a subject with a given theta value. The expected curves are produced as the summation of the expected probabilities of being in each of the categories over all the items at different theta locations. The vertical axis of the empirical category characteristic curves reflects the actual frequency of being in each of the categories for a subject with a given theta value. After putting the theta scale into small intervals, the empirical curves are produced as the actual frequency of choosing each of the categories averaged across examinees in different theta intervals. Since the patterns within grade turned out to be similar, forms within grades were averaged. Comparing the expected category curves and the empirical category curves, we can see that the peak and the intersection of different categories appear at similar theta locations in both plots. The expected plot and its empirical counterpart resemble each other to a certain extent. The bumps on the empirical curves are anticipated, which is due to the insufficient sample. In general, the expected and the empirical category characteristic curves suggest that the graded response model is a good representation of the real data.

Comparing the category curves across grades, we can see that the theta location where the slow-correct is the most frequent option shifts to the left of the scale as grade goes up. This suggests that the ability interval where the most frequent response is answering an item correct but slowly is lower in 5^{th} grade than in 3^{rd} grade. The empirical

curves of fast-correct and slow-correct overlap at the lower end of the theta scale in 3rd grade and 4th grade, and it is because there are insufficient students at these theta intervals. Other than that, the category curves have similar characteristics across grade.

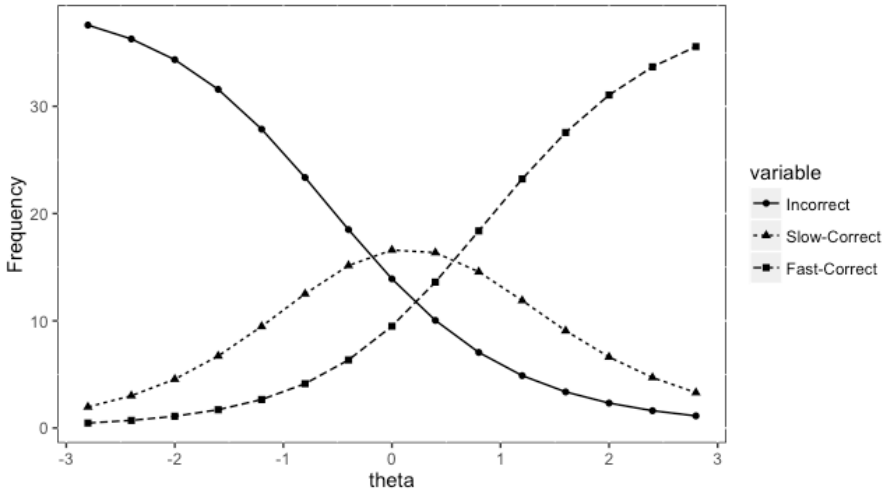

Figure 8.1. Expected category characteristic curves of Grade 3 forms averaged for the graded response model

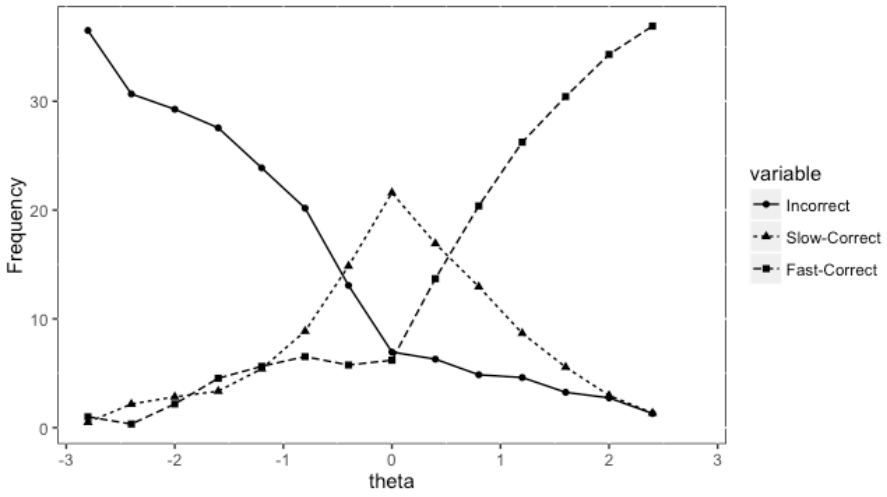

Figure 8.2. Empirical category characteristic curves of Grade 3 forms averaged for the graded response model

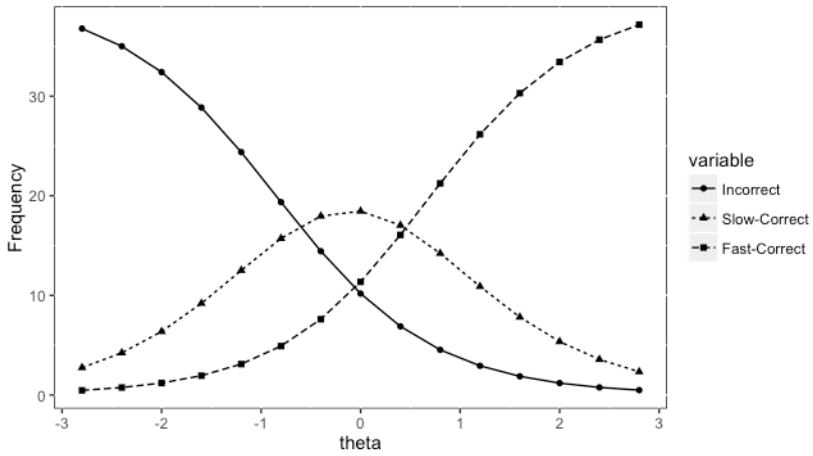

Figure 8.3. Expected category characteristic curves of Grade 4 forms averaged for the graded response model

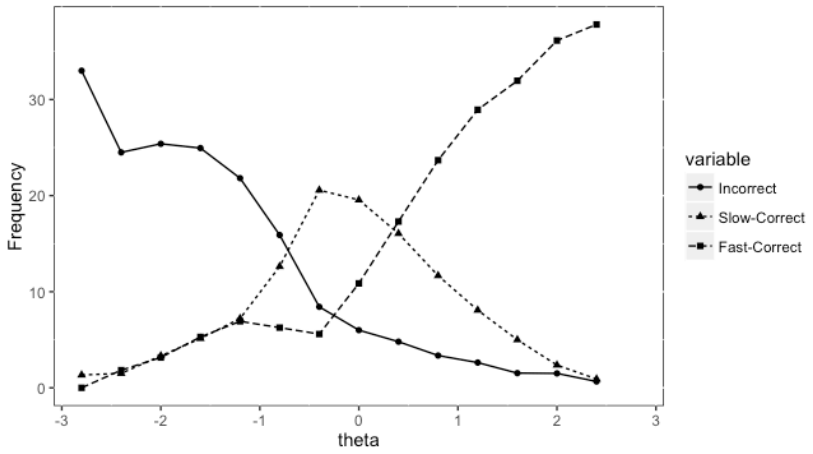

Figure 8.4. Empirical category characteristic curves of Grade 4 forms averaged for the graded response model

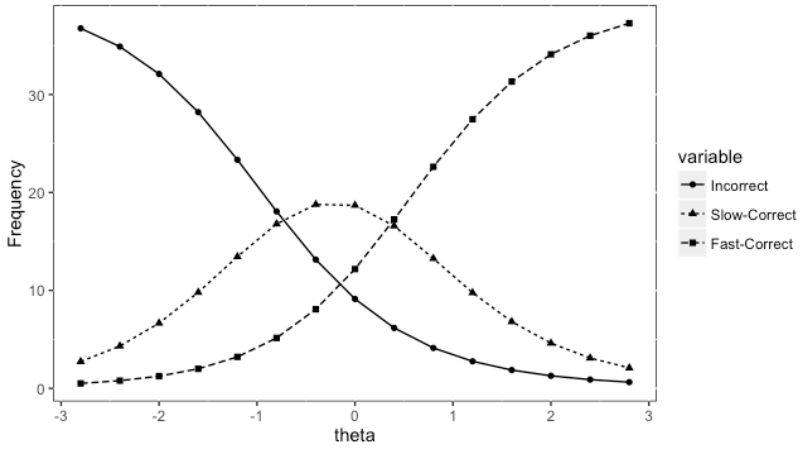

Figure 8.5. Expected category characteristic curves of Grade 5 forms averaged for the graded response model

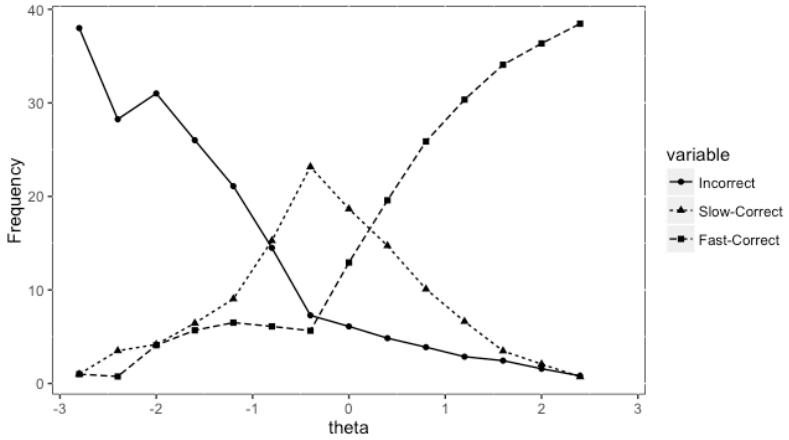

Figure 8.6. Empirical category characteristic curves of Grade 5 forms averaged for the graded response model

In general, the polytomously scored models have higher reliability than when the item is scored dichotomously. The graded scoring provides the best reliability and model fit. In terms of the information function, polytomously scored models tend to provide

greater information toward the upper end of the continuum than when item is dichotomously scored. The graded scoring results in greater information than other polytomous scoring models, particularly above the theta location of 0; moreover, the information is over a broader range of theta. The results suggest that combining the response times provides greater information.

Conditional Models

To further investigate the latent traits underlying the fast-correct and slow-correct responses, the next step was to utilize the RCIRT model, where the original response variables were scored into sets of parent variables (X_{kj}) for accuracy and sets of child variables ($X_{j\prime}$) for efficiency. The correctness of the response is represented by a parent item: $X_{kj} = 1$ if correct, $X_{kj} = 0$ if incorrect. Conditional on the parent item, a child item represents the speed of the response: $X_{j\prime} = 1$ if fast and $X_{kj} = 1$, $X_{j\prime} = 0$ if slow and $X_{kj} = 1$, $X_{j\prime} =$ missing if $X_{kj} = 0$. With the parent variables and child variables, these models are built and compared with each other: the unidimensional 2PL model with the parent variables only, which we call the *2PL Parent* model; the unidimensional 2PL model with the child variables only, which we call the *2PL Child* model; the unidimensional 2PL model with the parent variables and child variables combined, which we call the *2PL Parent-Child* model; and the multidimensional 2PL model of a simple structure with the parent variables and child variables combined, which we call the *MIRT* model.

Table 9 shows the theoretical marginal reliabilities of unidimensional models for each form and for each grade. The unidimensional 2PL model of the combined parent

and child variables has higher marginal reliabilities than the 2PL models with the parent variables only or the child variables only. When comparing the marginal reliabilities across grade, we can see that the marginal reliability of the 2PL Parent model decreases as the grade goes up, whereas the marginal reliabilities of the 2PL Child model and the 2PL Parent-Child model increase as the grade goes up. The increase of the marginal reliability of the models involving child variables is reasonable, since students in higher grades tend to have more accurate responses and therefore have less missing data on the child variables. This suggests the advantage of the RCIRT model since it is able to capture the information reflected by changes of grade.

Table 9

Marginal Reliabilities of 2PL Models across Form and across Grade

Form	2PL Parent	2PL Child	2PL Parent-Child
3.1	0.912	0.928	0.942
3.2	0.907	0.926	0.937
3.3	0.902	0.926	0.938
Grade 3	0.907	0.926	0.939
4.1	0.892	0.927	0.943
4.2	0.893	0.925	0.945
4.3	0.894	0.926	0.937
Grade 4	0.893	0.926	0.942
5.1	0.887	0.926	0.943
5.2	0.874	0.927	0.948
5.3	0.883	0.927	0.950
Grade 5	0.881	0.927	0.947

Table 10

Empirical Reliabilities of 2PL Models and Subscore Reliabilities of MIRT Model across Form and across Grade

Form	2PL Parent	2PL Child	2PL Parent -Child	MIRT Parent	MIRT Child
3.1	0.904	0.870	0.936	0.904	0.869
3.2	0.897	0.871	0.931	0.897	0.871
3.3	0.891	0.876	0.930	0.891	0.876
Grade 3	0.897	0.872	0.932	0.897	0.872
4.1	0.887	0.885	0.939	0.887	0.885
4.2	0.887	0.878	0.938	0.888	0.879
4.3	0.886	0.879	0.931	0.886	0.879
Grade 4	0.887	0.881	0.936	0.887	0.881
5.1	0.887	0.892	0.938	0.887	0.892
5.2	0.874	0.897	0.942	0.876	0.897
5.3	0.885	0.895	0.946	0.885	0.895
Grade 5	0.882	0.894	0.942	0.883	0.895

Table 10 above shows the empirical marginal reliabilities of the undimensional 2PL models and the empirical subscore reliabilities of the MIRT model for each form and for each grade. The unidimensional 2PL model of the combined parent and child variables has higher empirical reliability than the 2PL models with the parent variables only or the child variables only. The subscore reliabilities of the MIRT model are almost the same as their counterparts of the unidimensional 2PL models. This suggests a low correlation between the two dimensions, since subscore reliability of the MIRT is a function of the correlation between dimensions. Comparing the empirical reliabilities across grade, we can see that the empirical reliability of the 2PL Parent model and the subscore reliability of the accuracy dimension decrease as the grade goes up; while as the

grade goes up, the empirical reliabilities of the 2PL Child model and the 2PL Parent-Child model increase and the subscore reliability of the efficiency dimension also increases. As discussed above, the increase of the empirical reliability of the models with child variables is reasonable, since in higher grades students tend to have more accurate responses (as shown in Table 2) and therefore have less missing data on the child variables.

Table 11

Correlation of Thetas between Accuracy and Efficiency of 2PL Models and MIRT Model

Form	2PL Parent vs 2PL Child	MIRT Parent vs MIRT Child (Sample)	MIRT Parent vs MIRT Child (Population)
3.1	0.01	0.02	0.06
3.2	-0.10	-0.12	-0.06
3.3	-0.10	-0.12	-0.06
4.1	-0.01	0.01	0.06
4.2	0.08	0.12	0.15
4.3	-0.03	-0.03	0.00
5.1	0.07	0.10	0.11
5.2	0.12	0.16	0.17
5.3	0.07	0.10	0.13

Table 11 above shows the sample correlation between the theta estimates of accuracy and the theta estimates of efficiency for the unidimensional models and for the MIRT model, and the population correlation between the accuracy dimension and the efficiency dimension for the MIRT model. The correlation between accuracy (parent) and efficiency (child) is very low, which is in agreement with the similarity of reliabilities of the unidimensional model and the multidimensional model. The low correlation is

consistent with findings of Partchev et al.'s (2013) study. No significant difference is detected of the correlation between accuracy and efficiency for different models.

Table 12 below reports the test information functions averaged across the theta scale for the 2PL model with the parent variables only, the 2PL model with the child variables only, the 2PL model with the combined parent and child variables, and the MIRT model. It also reports the expected information functions for the child variables averaged across the theta scale. We can see that the information functions of models with the combined parent and child variables are higher than the models with only the parent variables or only the child variables.

Table 12

Average Information of 2PL Models and MIRT Model

Form	2PL Parent	2PL Child	Expected Child	2PL Parent-Child	MIRT Parent-Child
3.1	10.45	12.85	8.03	14.72	11.57
3.2	9.23	13.42	8.17	13.24	11.35
3.3	10.43	13.45	8.87	14.34	11.89
4.1	10.81	11.73	8.35	16.95	11.26
4.2	10.67	11.08	7.86	16.49	10.95
4.3	10.92	12.05	8.52	15.64	11.40
5.1	10.18	11.02	7.94	15.37	10.52
5.2	11.07	12.76	9.49	17.54	11.77
5.3	11.56	12.48	9.12	18.60	11.94

Table 13

AIC, BIC, -2LL and RMSEA of Conditional Models of Different Forms

Form		AIC	BIC	-2LL	RMSEA
3.1	2PL Parent	20678.46	21022.08	20518.46	0.03
	2PL Child	12646.18	12989.65	12486.18	0.02
	2PL Parent-Child	36663.53	37350.48	36343.54	0.06
	MIRT Parent-Child	33321.31	34012.55	32999.30	0.02
3.2	2PL Parent	20994.03	21334.34	20834.04	0.03
	2PL Child	11702.37	12042.68	11542.37	0.04
	2PL Parent-Child	36347.73	37028.35	36027.74	0.07
	MIRT Parent-Child	32694.20	33379.07	32372.20	0.03
3.3	2PL Parent	19173.22	19513.83	19013.22	0.03
	2PL Child	12418.44	12759.05	12258.44	0.03
	2PL Parent-Child	35468.27	36149.49	35148.26	0.07
	MIRT Parent-Child	31589.34	32274.83	15633.67	0.03
4.1	2PL Parent	18374.25	18714.71	18214.25	0.03
	2PL Child	14682.41	15022.87	14522.41	0.05
	2PL Parent-Child	36251.67	36932.59	35931.68	0.07
	MIRT Parent-Child	33061.39	33746.57	32739.38	0.04
4.2	2PL Parent	17626.00	17963.17	17466.00	0.03
	2PL Child	14218.74	14555.91	14058.74	0.04
	2PL Parent-Child	34622.89	35297.23	34302.90	0.06
	MIRT Parent-Child	31843.73	32522.28	31521.72	0.03
4.3	2PL Parent	16754.50	17088.23	16594.50	0.03
	2PL Child	13124.29	13457.86	12964.29	0.04
	2PL Parent-Child	32988.45	33655.59	32668.46	0.07
	MIRT Parent-Child	29872.07	30543.38	29550.08	0.03
5.1	2PL Parent	17042.43	17372.06	16882.43	0.03
	2PL Child	13430.54	13760.17	13270.54	0.04
	2PL Parent-Child	33118.37	33777.62	32798.36	0.07
	MIRT Parent-Child	30476.18	31139.54	30154.18	0.03
5.2	2PL Parent	14187.41	14509.87	14027.41	0.04
	2PL Child	12211.77	12534.23	12051.77	0.05
	2PL Parent-Child	28903.53	29548.44	28583.54	0.07
	MIRT Parent-Child	26400.13	27049.07	26078.12	0.04
5.3	2PL Parent	13899.36	14219.27	13739.36	0.03
	2PL Child	11403.83	11723.74	11243.83	0.04
	2PL Parent-Child	27908.02	28547.85	27588.02	0.07
	MIRT Parent-Child	25306.66	25950.48	24984.66	0.04

The model fit statistics for the unidimensional and multidimensional models are presented in Table 13 above. Note that we could only compare the relative fit statistics (AIC, BIC, and -2LL) of the undimensional 2PL model and MIRT 2PL model with the combined parent and child variables. The MIRT model has smaller AIC, BIC and -2LL than the unidimensional 2PL model for each form. In terms of the absolute fit statistics (RMSEA), the MIRT 2PL model has smaller RMSEA than other models.

Figures 9.1-9.9 show the test information functions of the 2PL model with the parent variables only, the 2PL model with the child variables only, and the expected information functions of the child variables. The expected information function of the child variables resembles the shape of the information function of the 2PL child model, and they all center around the theta location of 0.

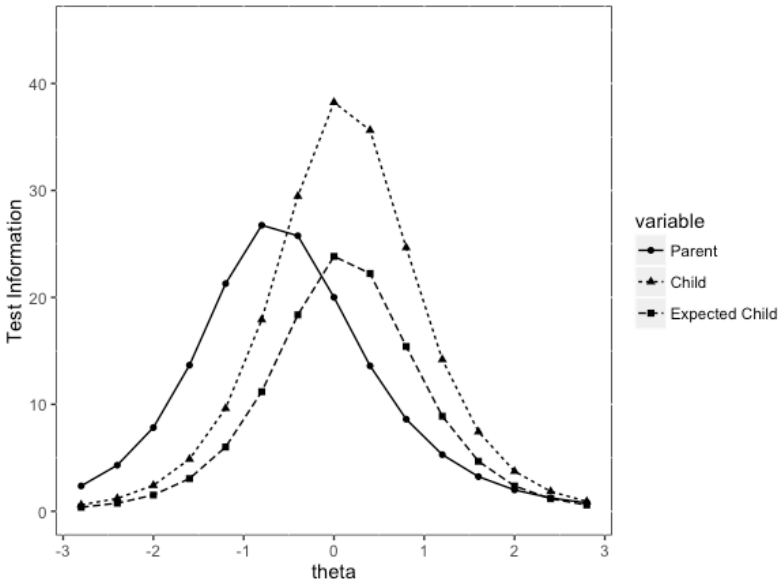

Figure 9.1. Parent information, child information and expected child information functions for Form 3.1

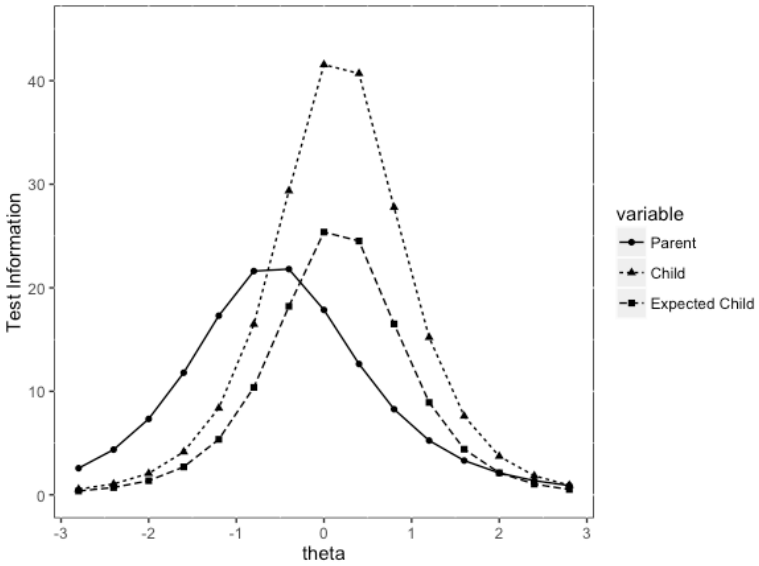

Figure 9.2. Parent information, child information and expected child information functions for Form 3.2

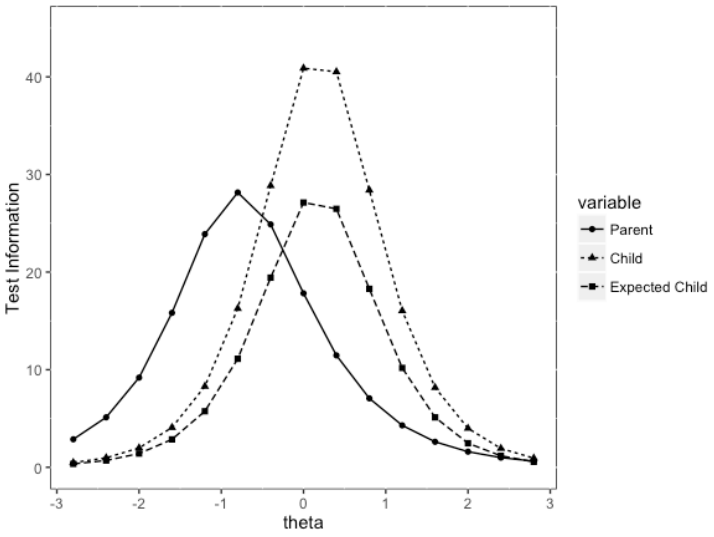

Figure 9.3. Parent information, child information and expected child information functions for Form 3.3

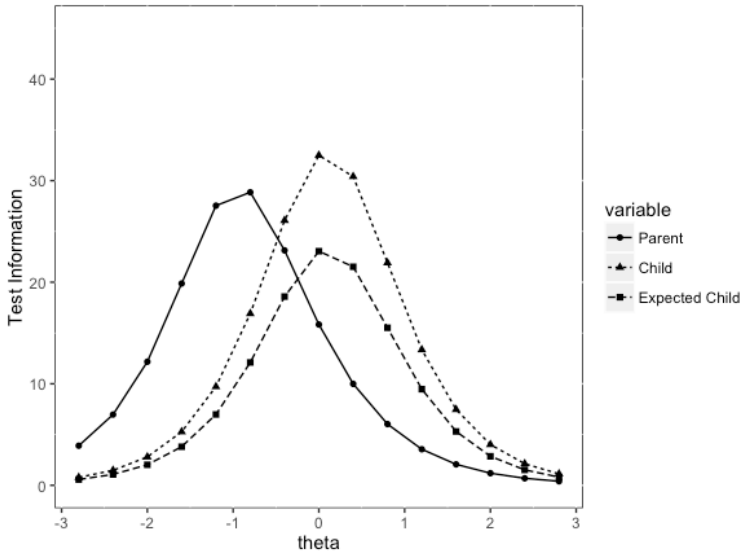

Figure 9.4. Parent information, child information and expected child information functions for Form 4.1

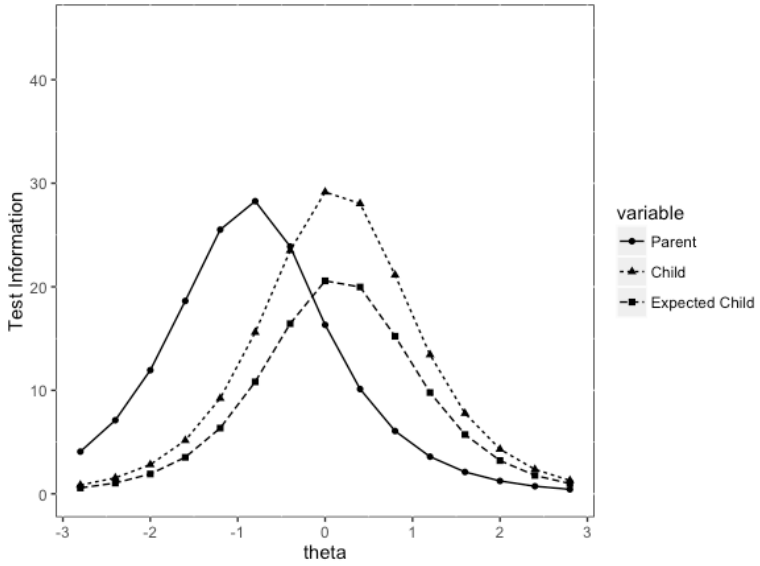

Figure 9.5. Parent information, child information and expected child information functions for Form 4.2

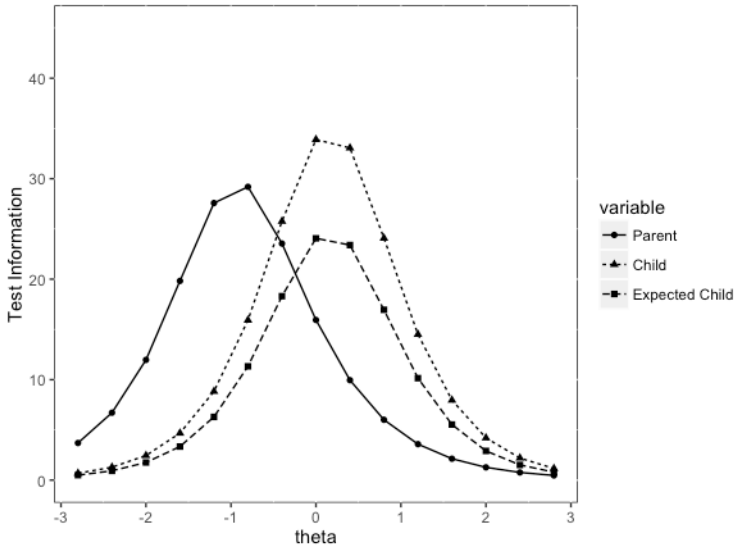

Figure 9.6. Parent information, child information and expected child information functions for Form 4.3

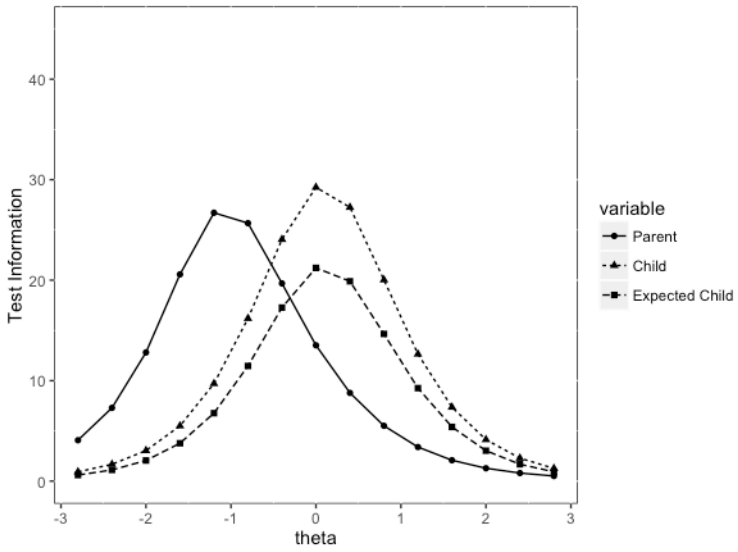

Figure 9.7. Parent information, child information and expected child information functions for Form 5.1

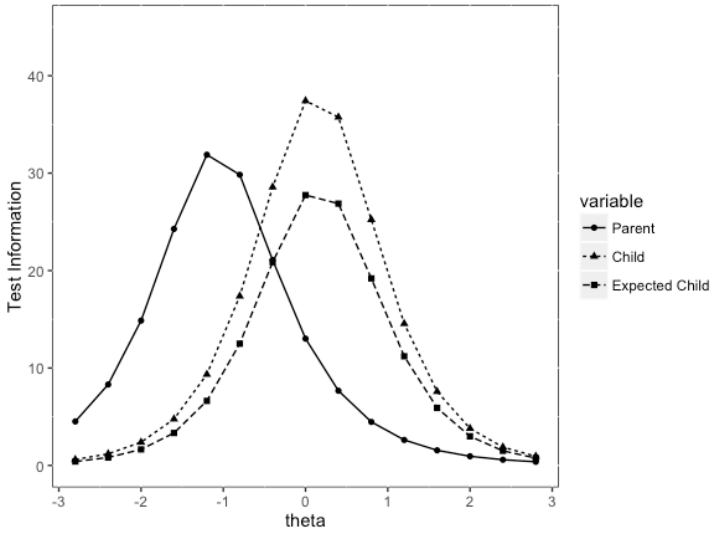

Figure 9.8. Parent information, child information and expected child information functions for Form 5.2

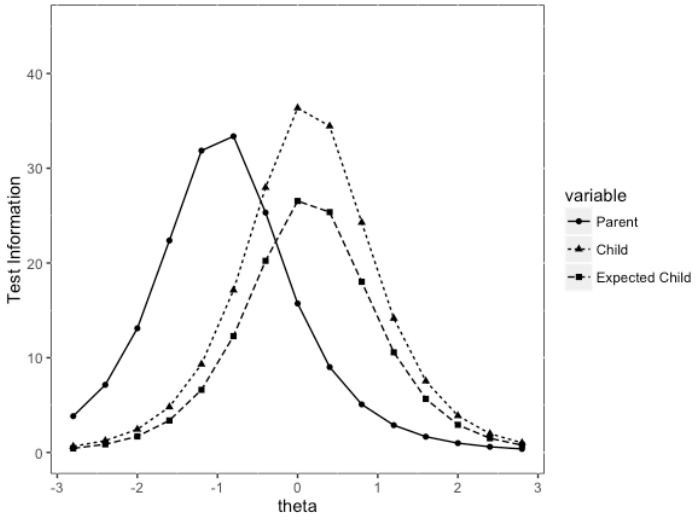

Figure 9.9. Parent information, child information and expected child information functions for Form 5.3

Table 14 reports the conditional standard error averaged across examinees at different theta intervals. The two models with the combined parent variables and child variables provide lower standard errors than the models with the parent variables only or with the child variables only. In addition, the unidimensional model with the combined parent and child variables provides lower standard errors than the multidimensional model with the combined variables.

Table 14

Average Conditional Standard Error at Different Theta Intervals for Each Form

Form	Model	Below -2.5	[-2.5, -1.5]	[-1.5, -0.5]	[-0.5, 0.5]	[0.5, 1.5]	[1.5, 2.5]	Above 2.5
3.1	2PL Parent	NA	0.319	0.215	0.253	0.379	0.518	NA
	2PL Child	NA	0.519	0.396	0.252	0.328	0.488	NA
	2PL Parent-Child	NA	0.288	0.194	0.215	0.297	0.396	0.511
	MIRT Parent	NA	0.319	0.217	0.254	0.381	0.521	NA
	MIRT Child	NA	0.517	0.396	0.253	0.329	0.484	NA
3.2	2PL Parent	NA	0.323	0.235	0.266	0.388	0.529	NA
	2PL Child	NA	0.545	0.415	0.253	0.295	0.466	NA
	2PL Parent-Child	NA	0.366	0.205	0.223	0.308	0.430	0.527
	MIRT Parent	NA	0.324	0.236	0.267	0.389	0.525	NA
	MIRT Child	NA	0.541	0.403	0.249	0.301	0.473	NA
3.3	2PL Parent	NA	0.348	0.215	0.269	0.419	0.553	NA
	2PL Child	NA	0.538	0.419	0.242	0.285	0.484	NA
	2PL Parent-Child	NA	0.373	0.196	0.234	0.307	0.422	0.522
	MIRT Parent	NA	0.326	0.215	0.268	0.420	0.553	NA
	MIRT Child	NA	0.531	0.415	0.242	0.289	0.498	NA
4.1	2PL Parent	NA	0.273	0.203	0.274	0.438	0.581	NA

	2PL Child	NA	0.528	0.351	0.246	0.309	0.463	NA
	2PL Parent-Child	NA	0.271	0.170	0.213	0.298	0.407	0.497
	MIRT Parent	NA	0.257	0.205	0.277	0.442	0.586	NA
	MIRT Child	NA	0.523	0.349	0.242	0.317	0.468	NA
4.2	2PL Parent	0.432	0.290	0.208	0.282	0.430	0.584	NA
	2PL Child	NA	0.532	0.366	0.269	0.302	0.435	NA
	2PL Parent-Child	NA	0.387	0.183	0.214	0.289	0.395	0.487
	MIRT Parent	0.433	0.283	0.208	0.280	0.433	0.589	NA
	MIRT Child	NA	0.521	0.376	0.262	0.298	0.430	NA
4.3	2PL Parent	NA	0.268	0.209	0.270	0.441	0.585	NA
	2PL Child	NA	0.538	0.394	0.240	0.294	0.455	NA
	2PL Parent-Child	NA	0.277	0.181	0.224	0.319	0.444	0.541
	MIRT Parent	NA	0.260	0.211	0.271	0.444	0.586	NA
	MIRT Child	NA	0.539	0.392	0.239	0.296	0.468	NA
5.1	2PL Parent	NA	0.230	0.206	0.281	0.427	0.582	NA
	2PL Child	NA	0.490	0.350	0.250	0.297	0.448	NA
	2PL Parent-Child	NA	0.230	0.195	0.219	0.283	0.375	0.474
	MIRT Parent	NA	0.226	0.213	0.288	0.438	0.584	NA
	MIRT Child	NA	0.479	0.339	0.249	0.309	0.449	NA
5.2	2PL Parent	0.359	0.214	0.197	0.291	0.468	0.604	NA
	2PL Child	NA	0.521	0.367	0.220	0.269	0.458	NA
	2PL Parent-Child	NA	0.285	0.179	0.212	0.281	0.382	0.501
	MIRT Parent	0.357	0.203	0.204	0.303	0.481	0.617	NA
	MIRT Child	NA	0.503	0.358	0.216	0.274	0.471	NA
5.3	2PL Parent	NA	0.235	0.189	0.266	0.459	0.604	NA
	2PL Child	NA	0.514	0.349	0.230	0.282	0.401	NA
	2PL Parent-	NA	0.222	0.166	0.197	0.275	0.380	0.478

Child							
MIRT Parent	NA	0.220	0.193	0.276	0.467	0.611	NA
MIRT Child	NA	0.505	0.338	0.228	0.289	0.410	NA

Note. NA means the value is not available since there are no cases in the interval.

Connecting Polytomous Models and Conditional Models

In summary, models with the combined parent variables and child variables provide higher reliability, higher information function, and lower conditional standard error than models with parent variables only and with child variables only. With the parent variables and child variables combined, the multidimensional models have better fit, but lower reliability, lower information function and higher conditional standard error than the unidimensional models. Therefore, it is reasonable to suggest that the parent variables and child variables contribute one dimension.

To further investigate the latent traits underlying the polytomous and conditional models, I looked at the correlation of the theta estimates obtained from the polytomous model –the graded response model was chosen based on its advantage – and those obtained from the conditional models. Table 15 shows the correlation of the theta estimates of the graded response model against those of the 2PL Parent model, 2PL Child model, 2PL Combined model, and two dimensions of MIRT model, for each form.

Table 15

Correlation of Theta estimates of the GRM against Those of the Conditional Models

Form	With: 2PL Parent	With: 2PL Child	With: 2PL Parent-Child	With: MIRT Parent	With: MIRT Child
3.1	0.88	0.36	0.98	0.88	0.37
3.2	0.86	0.31	0.98	0.86	0.30
3.3	0.83	0.34	0.98	0.83	0.33
4.1	0.84	0.43	0.98	0.84	0.43
4.2	0.86	0.49	0.99	0.87	0.51
4.3	0.83	0.41	0.97	0.83	0.41
5.1	0.82	0.53	0.99	0.83	0.54
5.2	0.81	0.58	0.99	0.83	0.59
5.3	0.82	0.51	0.99	0.83	0.52

We have already discussed the correlation of the theta estimates of the graded response model and the 2PL Parent model measuring accuracy and the correlation between different conditional models, therefore our next focus is on the correlation of the graded response model with conditional models involving child variables. As shown in Table 15, the correlation between the theta estimates of the graded response model with those of the unidimensional and multidimensional models measuring accuracy is relatively high, and it displays a decreasing trend as grade goes up. One the contrary, the correlation between the theta estimates of the graded response model with those of the unidimensional and multidimensional models measuring efficiency is low to moderate, and it displays an increasing trend as grade goes up. The trend across grade is consistent with what we observed with the reliabilities of the corresponding models.

Mostly importantly, the theta estimates of the graded response model across different forms are highly correlated with those of the 2PL model with the combined parent variables and child variable. It suggests these two models measure a similar latent

trait. Given that the graded response model measures a general dimension, which is the reading comprehension fluency according to the theory, we suggest that the conditional model with the combined parent variables and child variables contributes to a single dimension and it is reading comprehension fluency as measured in the polytomous model.

The correlation of the polytomous model and the conditional models implies that they measure the same trait but in different ways. Polytomously scored variables incorporate response times with response accuracy and model them as a general dimension of reading comprehension fluency. The parent variables measuring ability in the conditional models denote the transition of incorrect to correct, and the child variable measuring efficiency in the conditional models denote the transition of slow-correct to fast-correct, and together they contribute a single dimension which is the reading comprehension fluency. Both approaches, the polytomous scoring and the RCIRT (unidimensional or multidimensional), exhibit advantages over the traditional models where only the accuracy data are considered. The results indicate the importance of incorporating the response times into the estimation of person trait. The results correspond to theory of reading comprehension fluency and thus serve as evidence of construct validity of measuring reading comprehension fluency.

Chapter V: Conclusion and Discussion

Summary

The current study integrates response time information into an IRT framework in the context of a reading comprehension assessment. It aims at developing models applicable to a new instrument of reading comprehension to evaluate examinees' reading comprehension fluency. Chapter 1 and 2 stated the importance of recording and studying response times along with response accuracy. With online assessment becoming mainstream and collection of response time data becoming straightforward, there is an increasing volume of literature proposing approaches utilizing response times. Besides response accuracy, response time is another important measure in some intelligence tests. Response time has been used to measure constructs, to improve criterion-related validity and to evaluate abnormal behaviors. Chapter 1 states the research questions to be addressed in the current study: whether incorporating response times improves the construct validity of measuring the defined construct reading comprehension fluency, along with a better interpretation of the construct; and whether the proposed models improve the estimation of person trait parameters.

Chapter 2 summarizes approaches to model response times and response accuracy, different types of models incorporating response times, and candidate distributions of response times. There exist various interesting models for response times, but there is not much agreement on which models to use. Three approaches have been advocated in previous studies, to model response times exclusively (e.g., Baayen et al., 2008; Rouder

et al., 2003; Scheiblechner, 1979), or concurrently but separately from response accuracy (e.g., Gorin, 2005; Mulholland et al., 1980), or concurrently and jointly with response accuracy (e.g., Klein Entink et al., 2009; Loeys et al., 2011; Roskam, 1997; van der Linden, 1999; 2007). To model response times, previous studies have adapted IRT models (e.g., Roskam, 1997; Wang & Hanson, 2005), mixed models (e.g., van Breukelen, 2005), hierarchical structures (e.g., van der Linden, 2007; Loeys et al., 2011), Cox proportional Hazards models (e.g., Ranger & Ortner, 2011; Wang et al., 2013), mixture models (e.g., Schnipke & Scrams, 1997; Wang & Xu, 2015) and so on. Also there is great variety in the distributions that have been assumed when modeling response times: log normal (e.g., Thissen, 1983; van der Linden, 2007), exponential (e.g., Scheiblechner, 1979), gamma (e.g., Maris, 1993; Verhelst et al., 1997), Weibull (e.g., Rouder et al., 2003; Loeys et al., 2011), and the Box-Cox normal model (Klein Entink et al., 2009). The variety of modeling approaches brings complexity to response time modeling

Following the conversation of models in previous studies, Chapter 2 discusses the latent trait underlying response times and the relationship between speed (measured by response times) and ability (measured by accuracy). A construct named reading comprehension fluency is introduced as a product of comprehension ability and speed. To understand the construct of reading comprehension fluency and to avoid the complexity of choosing from existing models, the current study dichotomized the response times and modeled response times and responses jointly with proposed IRT based models: polytomous models and RCIRT models. Chapter 2 further discusses the theoretical background of the polytomous model and the RCIRT model. One notable feature that

distinguishes both approaches from previous studies is that they utilize only *correct* response times.

As illustrated in Chapter 3, data from an online assessment of reading comprehension named MOCCA were collected to study reading comprehension fluency for students in grade 3-5. Chapter 3 introduces the features of this new instrument of reading comprehension and evaluates the unidimensional polytomous models (PCM, GPCM and GRM) and RCIRT models, based on dependent variables including model fit statistics, marginal reliability, information functions, characteristic curves and average conditional standard errors.

As presented in Chapter 4, the results for the polytomous models indicate that compared with the dichotomously scored model with only accuracy variables, the polytomously scored model involving response time information is preferred due to its advantages in reliability, model fit, and information function. The polytomous models fit well and provide more information with most of the additional information at the *top* end of the scale. Among them, the GRM is the most reliable and has the best model fit; it also has the highest information function, especially for students with higher levels of reading comprehension fluency. After summing category response curves across items, the overall category response curve identifies the intervals along the comprehension continuum in which correct responses tend to be more commonly efficient or inefficient.

The results for the conditional models suggest that the RCIRT model obtained from both parent variables and child variables has higher reliability, higher information function, and lower conditional standard error than models with either solely parent

variables or solely child variables. Also, the results imply a unidimensional model works well for the parent variables and child variables combined, thus it is reasonable to suggest that response accuracy and response times together contribute to the measure of one latent trait, which is reading comprehension fluency according to definition.

Discussion

Response time and response accuracy work together as measures of reading comprehension fluency. The theoretical framework of the construct evaluated in the current study is shown in Figure 10.

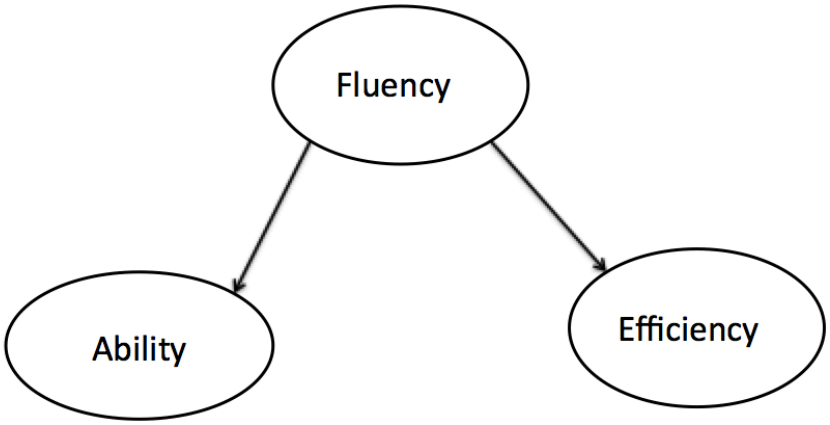

Figure 10. Theoretical framework of the construct of reading comprehension fluency

The model with polytomously scored variables provides the measure of a general construct of reading comprehension fluency, denoted as θ^*, and the RCIRT models provide specific measures of two essential components of reading comprehension fluency: comprehension ability (θ_1) and comprehension efficiency (θ_2). The comprehension ability determines the transition of incorrect responses to correct responses and is measured by response accuracy, and we have $\theta_1 = \lambda_1\theta^*$, where λ_1

denotes the correlation between the estimates of reading comprehension fluency and comprehension ability. The comprehension efficiency determines the transition of slow-correct responses to fast-correct responses and is measured by the response times conditional on accuracy, and $\theta_2 = \lambda_2\theta^*$, where λ_2 denotes the correlation between the estimates of reading comprehension fluency and comprehension efficiency. It is important to note that Figure 1 is not a structural equation modeling (SEM) framework, since the latent trait of reading comprehension fluency (θ^*) is directly measured by the polytomous model. Moreover, the current study shows that the efficiency estimates obtained from the child variables and the accuracy estimates from the parent variables are not directly correlated, which is consistent with Partchev et al.'s (2013) study that speed and ability were more or less uncorrelated in a verbal analogy test. The interpretation of the transitions of reading comprehension fluency serves as an evidence of construct validity for measuring the construct of reading comprehension fluency as defined in theory.

Even though the current study is similar to Partchev and De Boeck's study (2012; Partchev et al., 2013) in terms of dichotomizing response times and using RCIRT models, the two studies are fundamentally different from each other. As shown in Figure 11, Partchev and De Boeck's (2012) model proposes a general construct of intelligence and further differentiates fast intelligence and slow intelligence. The fast intelligence determines the transition of fast-and-incorrect response to fast-and-correct response, and the slow intelligence determines the transition of slow-and-incorrect response to slow-and-correct response. The current study proposes a general construct of reading

comprehension fluency and its two essential components are comprehension ability and comprehension efficiency. The ability decides the transition of incorrect response to correct response, and the efficiency decides the transition of slow-correct response to fast-correct response.

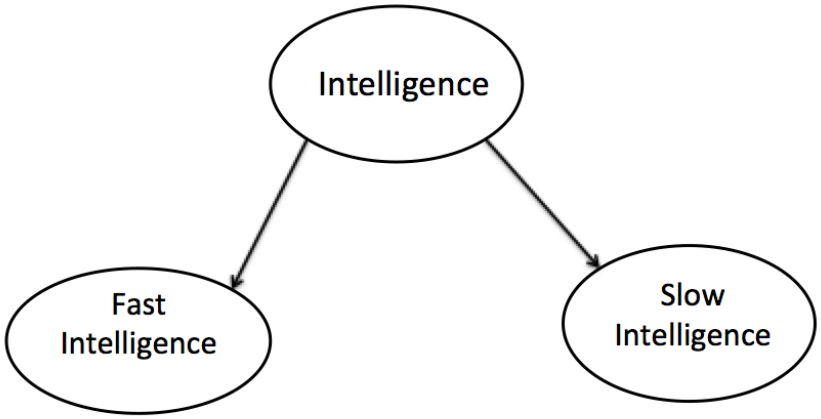

Figure 11. Theoretical framework of Partchev and De Boeck's (2012) fast intelligence and slow intelligence

Limitation and Future Work

The MOCCA test is an ongoing test development project, and discussion of the limitations could help improve the instrument. This section discusses three limitations of the current study and proposes possible approaches addressing the limitations in future work. First, there are no linking items in forms across grade. Without the linking items, it is not possible to equate the theta estimates of different forms and precisely describe the growth trend of reading comprehension fluency across grade. Therefore, the current study provides no evidence to evaluate the theory of development of reading comprehension

fluency across age or grade. In the next step of the test development, linking items will be developed and included in each of the forms.

Second, the missing values in the RCIRT models should be dealt with carefully. The child variables can be missing under some conditions. In the current study, the child variables are missing when the parent variables are incorrect responses. The missingness of the child variable is not missing completely at random (MCAR) because it is completely determined by the observed parent variable. However, it is missing at random (MAR) because it is solely a function of the observed parent variable, and then it can be ignorable in a limited sense (Mislevy & Wu, 1996). A large amount of missingness can give rise to sample size issues and large standard errors on the parameter estimates. As presented in Chapter 3, the missingness is not a big concern in the current study. In a more general case, missingness in the child variable could be a challenge. It may be possible to address this missing data problem with some form of imputation. In the future work, simulation research is needed to evaluate imputation methods for their effectiveness in imputing response variables modeled with an RCIRT model.

Last, unbiased external criterion variables can be obtained to provide evidences of criterion-related validity of measuring reading comprehension fluency using the MOCCA test. For example, in practice, teacher evaluations and parent feedbacks can be collected to further validate the reading comprehension fluency measured in the current study.

Significance

The importance of response times as a measure of psychological constructs has

been recognized and the literature of modeling response times has been growing during the past few decades. Different from various existing psychometric models, the current study employs the idea of reading comprehension fluency in the literature of reading field and proposes some IRT based models combining response times and response accuracy. To better understand the construct of reading comprehension fluency, the current study evaluates reading comprehension fluency in two approaches: one with polytomously scored variables and one with conditional child variables.

The current study avoids the complexity of choosing existing psychometric models of response times and utilizes well-known IRT models. It extends the psychometric literature of modeling response times and the cognitive psychology literature in the reading field. Findings of the current study correspond to the theoretical framework of the construct of reading comprehension fluency, and lead to a better interpretation of the latent trait of reading comprehension fluency. It shows that response time information can be used to enhance the construct validity of measuring the construct of reading comprehension fluency. The models developed in the current study identify the intervals along the comprehension continuum in which the correct responses tend to be more commonly efficient or inefficient. In practice, these models can be used to provide teachers with accurate information about the students' level of reading comprehension fluency for purpose of instructional differentiation.

Reference

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Andrich, D. & Kreiner, S. (2010). Quantifying response dependency between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*, 181 – 192.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412.

Bassili, J. (1996). The how and why of resposne latency measurement in telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research*, (pp. 319–346). San Francisco: Jossey-Bass.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Carlson, S., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences, 32*, 40 – 53.

Chard, D. J., Vaughn, S., & Tyler, B. (2002). A synthesis of research on effective interventions for building reading comprehension fluency with elementary

students with learning disabilities. *Journal of Learning Disabilities, 35*(5), 386–406.

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis 1. *Educational and Psychological Measurement, 5*(4), 411-427.

Davison, M. L., Liu, B., Wang, Q., Su, S., Biancarosa, G., Carlson, S., & Seipel, B. (2016). Response Contingent Item Response Theory. *Unpublished manuscript*.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Dennis, I., & Evans, J. S. B. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology, 87*(1), 105-129.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83*(6), 1441.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*(4), 351-373.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Holden, R. & Kroner, D. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment, 4*, 170–173.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434-446.

Kahane, M., & Loftus, G. (1999). Response time versus accuracy in human memory. In R.J. Sternberg (Ed.). *The Nature of Cognition* (pp. 323–384). Cambridge (MA): MIT.

Kennedy, M. (1930). Speed as a personality trait. *Journal of Social Psychology, 1*, 286–298.

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*, 621-640.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika, 76*(3), 487-503.

Lohman, D. F. (1989). Estimating individual differences in information processing using speed-accuracy models. In Kanfer, R., Ackerman, P., & Cudeck, R. (Eds.), *Abilities, motivation, and methodology: The Minnesota symposium on learning and individual differences* (pp. 119-156). Lawrence Erlbaum Associates Hillsdale, NJ.

Luce, R.D. (1986). *Response times: Their roles in inferring elementary mental organization*. Oxford, UK: Oxford University Press.

Maris, E. (1993). Additive and Multiplicative Models for Gamma Distributed Random Variables, and Their Application as Psychometric Models for Response Times. *Psychometrika, 58*, 445-469.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 521–538.

Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new direction. *Annals of Dyslexia, 49*, 283–306.

Mislevy, R. & Wu, P. K. (1996). *Missing responses and IRT estimation: Omits, choice, time limits, and adaptive testing* (ETS RR-96-30-NR). Princeton NJ: Educational Testing Service.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate behavioral research, 50*(1), 56-74.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive psychology, 12*(2), 252-284.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated?. *Intelligence, 40*(1), 23-32.

Partchev, I., De Boeck, P., & Steyer, R. (2013). How much power and speed is measured in this test? *Assessment, 20*(2), 242-252.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: an illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing, 28*, 31 – 56.

Ranger, J., & Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika, 77*(1), 31-47.

Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology, 65*(2), 334-349.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review, 111*(2), 333.

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer.

Rouder, J., Sun, D., Speckman, P., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68*, 589-606.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72*(4), 621-642.

Rouder, J., Tuerlinckx, F., Speckman, P., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review, 15*(6), 1201–1208.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19*, 18–38.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human in- formation processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Siem, F. (1996). The use of response latencies to enhance self-report personality measures. *Military Psychology, 8*, 15–27.

Spearman, C. (1927). *The abilities of man*. New York, NY: Macmillan.

Tate, M. W. (1948). Individual difference in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement, 8*, 353–374.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J.Weiss(Eds), *New horizons in testing* (pp.179–203). New York: Academic Press.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple‐categorical‐response models. *Journal of Educational Measurement, 26*(3), 247-260.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates Hillsdale, NJ.

Thorndike, E.L., Bregman, E.O., Cobb, M.V., & Woodyard, E. (1926). *The measurement of intelligence*. New York: Teachers College Bureau of Publications.

Townsend, J. T, & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge, England: Cambridge University Press.

van Breukelen, G. J. (2005). Psychometric modelling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*,359–376.

van der Linden, W.J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*, 21–29.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*, 247-272.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365–384.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive Testing. *Applied Psychological Measurement, 23*(3), 195-210.

Verhelst, N. D., Verstralen, H. H., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In *Handbook of modern item response theory* (pp. 169-185). Springer New York.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*, 323-339.

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics, 38*(4), 381-417.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456-477.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort‐moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Tech. Rep. No. TR-10). Princeton, NJ: Educational Testing Service.

Appendix

Table A1.

Mean and Standard Deviation of Theta Estimates for the PCM, GPCM, GRM and 2PL Models

Form	PCM	GPCM	GRM	2PL
3.1	0.01 (0.96)	0.00 (0.96)	0.00 (0.97)	0.00 (0.95)
3.2	0.01 (0.95)	0.00 (0.96)	0.00 (0.96)	0.00 (0.95)
3.3	0.01 (0.95)	0.00 (0.96)	0.00 (0.96)	0.00 (0.95)
4.1	0.00 (0.96)	0.00 (0.96)	0.00 (0.97)	0.00 (0.94)
4.2	0.01 (0.96)	0.00 (0.97)	0.00 (0.97)	0.00 (0.94)
4.3	0.01 (0.96)	0.00 (0.96)	0.00 (0.96)	0.00 (0.94)
5.1	0.00 (0.96)	0.00 (0.97)	0.00 (0.97)	0.00 (0.94)
5.2	0.01 (0.97)	0.00 (0.97)	0.00 (0.97)	0.00 (0.94)
5.3	0.00 (0.97)	0.00 (0.97)	0.00 (0.97)	0.00 (0.94)

Note. Numbers in the brackets are standard deviations.